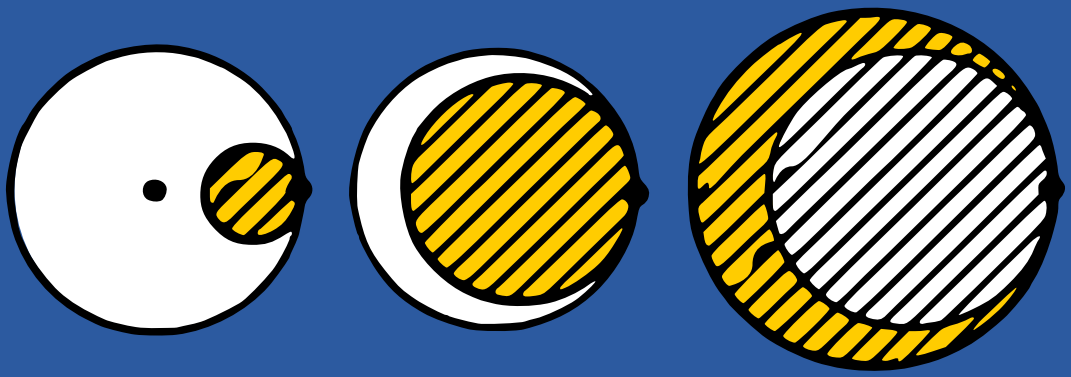
S. Godounov, V. Riabenki

SCHÉMAS AUX DIFFÉRENCES



Éditions Mir Moscou

С. К. ГОДУНОВ, В. С. РЯБЕНЬКИЙ

РАЗНОСТНЫЕ СХЕМЫ

Введение в теорию

ИЗДАТЕЛЬСТВО «НАУКА»
МОСКВА

S. GODOUNOV, V. RIABENKI

SCHÉMAS AUX DIFFÉRENCES

Introduction à la théorie

ÉDITIONS MIR · MOSCOU

Traduit du russe
par IRINA PÉTROVA

На французском языке

© Издательство «Наука», 1973

© Traduction française Editions Mir · 1977

TABLE DES MATIÈRES

Préface 11
Introduction 13

Première partie

ÉQUATIONS AUX DIFFÉRENCES ORDINAIRES

Chapitre premier. ÉQUATIONS AUX DIFFÉRENCES DU PREMIER ET DU SECOND ORDRE. EXEMPLES DE SCHÉMAS AUX DIFFÉRENCES 17

§1. Equations aux différences simples 17
1. Equations aux différences (17). 2. Ordre d'une équation aux différences (19). 3. Solution générale d'une équation aux différences (20).
Exercices 22

§2. Equation aux différences du premier ordre 22
1. Solution fondamentale (23). 2. Condition de borne de la solution fondamentale (24). 3. Solution particulière (24).
Exercices 26

§3. Equation aux différences du second ordre 26
1. Solution générale de l'équation homogène (27). 2. Solution générale de l'équation non homogène. Solution fondamentale (30). 3. Evaluation de la solution fondamentale par les coefficients de l'équation aux différences (34).
Exercices 35

Chapitre 2. PROBLÈME AUX LIMITES POUR UNE ÉQUATION DU SECOND ORDRE 38

§4. Position du problème. Critères de bon conditionnement 38
1. Position du problème (38). 2. Notion de bon conditionnement (39). 3. Critère suffisant de bon conditionnement (40). 4. Critère de bon conditionnement d'un problème aux limites à coefficients constants (41). 5. Critère de bon conditionnement d'un problème à coefficients variables (42). 6. Justification du critère de bon conditionnement d'un problème aux limites à coefficients constants (43). 7. Problèmes aux limites généraux pour des systèmes d'équations aux différences (47).
Exercices 49

§5. Balayage 50
1. Description (50). 2. Exemple d'algorithme numériquement instable (52).
Exercices 53

Chapitre 3. JUSTIFICATION DE LA MÉTHODE DU BALAYAGE 55

§6. Propriétés des problèmes aux limites bien conditionnés 55
1. Evaluation des solutions d'un problème aux limites à coefficients perturbés (55). 2. Démonstration du critère de bon conditionnement (58). 3. Propriétés des problèmes bien conditionnés (62).

§7. Justification de la méthode du balayage pour des problèmes aux limites bien conditionnés 63
1. Evaluation des coefficients de balayage (63). 2. Influence des erreurs d'arrondi sur le résultat (65).

Deuxième partie

SCHÉMAS AUX DIFFÉRENCES POUR LES ÉQUATIONS DIFFÉRENTIELLES ORDINAIRES

Chapitre 4. EXEMPLES ÉLÉMENTAIRES DE SCHÉMAS AUX DIFFÉRENCES 68

§8. Notions d'ordre de précision et d'approximation 68
1. Ordre de précision d'un schéma aux différences (68). 2. Vitesse de convergence de la solution d'une équation aux différences (71). 3. Ordre d'approximation (73).

§9. Un schéma aux différences instable 74
1. Procédés d'approximation d'une dérivée (74). 2. Exemple de schéma aux différences instable (74).

Chapitre 5. CONVERGENCE DES SOLUTIONS DES ÉQUATIONS AUX DIFFÉRENCES, CONSÉQUENCE DE L'APPROXIMATION ET DE LA STABILITÉ 77

§10. Convergence d'un schéma aux différences 77
1. Réseau et fonction définie sur un réseau (77). 2. Schémas aux différences convergents (82). 3. Vérification de la convergence d'un schéma aux différences (84).
Exercices 86

§11. Approximation d'un problème aux limites différentiel par un schéma aux différences 86
1. Résidu $\delta f^{(k)}$ (86). 2. Calcul du résidu (88). 3. Approximation de l'ordre de h^k (90). 4. Exemples (90). 5. Décomposition d'un schéma aux différences en sous-systèmes (93). 6. Remplacement des dérivées par des rapports aux différences (95). 7. D'autres méthodes de construction de schémas aux différences (97).
Exercice 97

§12. Définition de la stabilité d'un schéma aux différences. Convergence en tant que conséquence de l'approximation et de la stabilité 98
1. Définition de la stabilité (98). 2. Relation entre l'approximation, la stabilité et la convergence (100). 3. Schéma aux différences convergent pour une équation intégrale (105).

§13. Choix des normes 106

§14. Critère suffisant de stabilité des schémas aux différences pour le problème de Cauchy 113

1. Exemple préliminaire (113). 2. Schéma aux différences sous forme canonique (114). 3. Stabilité en tant que borne des normes des puissances de R_h (116). 4. Exemples d'étude de la stabilité (117). 5. Non-unicité de l'écriture canonique (122).
Exercices 123

§15. Critère spectral nécessaire de stabilité 124
1. Une condition nécessaire de stabilité est que les normes des puissances de l'opérateur R_h soient bornées (125). 2. Critère spectral de stabilité (126). 3. Discussion du critère spectral de stabilité (127).
Exercices 131

§16. Erreurs d'arrondi 131
1. Erreurs sur les coefficients (131). 2. Erreurs de calcul (134).

§17. Caractérisation quantitative de la stabilité 135

§18. Un procédé d'étude de la stabilité des problèmes non linéaires 141

Chapitre 6. SCHÉMAS AUX DIFFÉRENCES USUELS 143

§19. Schémas de Runge-Kutta et d'Adams 143
1. Schémas de Runge-Kutta (143). 2. Schémas d'Adams (145). 3. Remarques sur la stabilité (148). 4. Généralisation à des systèmes d'équations (149).

§20. Méthodes de résolution des problèmes aux limites 151
1. Méthode du tir (151). 2. Méthode du balayage (153). 3. Méthode de Newton (154).

Troisième partie

SCHÉMAS AUX DIFFÉRENCES POUR LES ÉQUATIONS AUX DÉRIVÉES PARTIELLES. NOTIONS FONDAMENTALES

Chapitre 7. PROCÉDÉS SIMPLES DE CONSTRUCTION ET D'ÉTUDE DE SCHÉMAS AUX DIFFÉRENCES 155

§21. Rappels et illustration des définitions principales 155
1. Convergence (155). 2. Approximation (156). 3. Stabilité (159).
Exercices 165

§22. Procédés simples de construction de schémas aux différences approximants 166
1. Remplacement des dérivées par des rapports aux différences (166). 2. Méthode des coefficients indéterminés (172). 3. Schémas de prédiction-correction (181).
Exercices 182

§23. Exemples de construction de conditions aux limites 183
Exercices 188

§24. Condition nécessaire de convergence de Courant, Friedrichs et Lewy 189
1. Condition de Courant, Friedrichs et Lewy (189). 2. Exemples de schémas aux différences pour le problème de Cauchy (190). 3. Exemples de schémas aux différences pour le problème de Dirichlet (195).
Exercices 197

Chapitre 8. CERTAINS GRANDS PROCÉDÉS D'ÉTUDE DE LA STABILITÉ 198

§25. Analyse spectrale du problème de Cauchy aux différences 198
1. Stabilité relativement aux données initiales (198). 2. Condition spectrale

nécessaire de stabilité (199). 3. Exemples (201). 4. Représentation intégrale d'une solution (208). 5. Lissage de la solution d'un problème aux différences en tant qu'effet de la viscosité d'approximation (212).
Exercices 214

§26. Principe des coefficients fixés 215
1. Fixation des coefficients aux points intérieurs (215). 2. Critère de Babenko et Guelfand (218).
Exercices 223

§27. Solutions de certains problèmes modèles sous forme des séries finies de Fourier 224
1. Séries de Fourier pour des fonctions définies sur un réseau (224). 2. Représentation des solutions de schémas aux différences pour l'équation de la chaleur sur un segment (227). 3. Représentation des solutions de schémas aux différences pour le problème plan de la chaleur (230). 4. Représentation de la solution d'un schéma aux différences pour le problème des cordes vibrantes (232).
Exercices 234

§28. Principe du maximum 235
1. Schéma aux différences explicite (235). 2. Schéma aux différences implicite (237). 3. Comparaison des schémas aux différences explicite et implicite (238).

Chapitre 9. LES SCHÉMAS AUX DIFFÉRENCES ET LE CALCUL DES SOLUTIONS GÉNÉRALISÉES 239

§29. Solution généralisée 239
1. Apparition des discontinuités (240). 2. Définition d'une solution généralisée (241). 3. Condition sur la ligne de discontinuité de la solution (242). 4. Suppression d'une discontinuité (244). 5. Autre définition d'une solution généralisée (245).

§30. Construction de schémas aux différences 246
1. Schéma avec viscosité artificielle (247). 2. Méthode des caractéristiques (247). 3. Schémas aux différences conservatifs (248).

Quatrième partie

PROBLÈMES À DEUX VARIABLES D'ESPACE

Chapitre 10. NOTION DE SCHÉMAS DE DÉCOMPOSITION AUX DIFFÉRENCES 254

§31. Construction de schémas de décomposition 254
Exercices 258
§32. Schémas aux différences économiques 258
Exercices 265

§33. Décomposition suivant les facteurs physiques 265

Chapitre 11. PROBLÈMES ELLIPTIQUES 266

§34. Un schéma aux différences très simple pour le problème de Dirichlet 266
1. Approximation (266). 2. Stabilité (267).
Exercices 270

§35. Méthode de stationnarisation 271
1. Idée de la méthode (271). 2. Analyse d'un schéma de stationnarisation explicite (273). 3. Schéma de directions alternées (276). 4. Choix de la

précision (277). 5. Limites d'applicabilité des méthodes (278).
Exercices 278

§36. Itérations à pas variable 279
1. Idée de Richardson (279). 2. Ensemble de paramètres de Tchebycheff (279). 3. Numérotation des paramètres d'itération. (282). 4. Méthode de Douglas-Rachford (285).
Exercices 288

§37. Méthode de Fédorenko 288
1. Idée de la méthode (289). 2. Description de l'algorithme (290).

Cinquième partie

STABILITÉ DES PROBLÈMES AUX LIMITES AUX DIFFÉRENCES ÉVOLUTIFS EN TANT QUE BORNE DES NORMES DES PUISSANCES D'UN OPÉRATEUR

Chapitre 12. CONSTRUCTION DE L'OPÉRATEUR DE PASSAGE 293

§38. Structure stratifiée des solutions des problèmes évolutifs 293
Exercices 296

§39. Ecriture des problèmes aux limites aux différences sous forme $u^{p+1} = R_h u^p + \tau \varrho^p$ 296
1. Forme canonique (296). 2. Stabilité en tant que borne uniforme des normes des puissances de l'opérateur R_h (299). 3. Exemple (302).
Exercices 304

§40. Utilisation des solutions particulières pour construire l'opérateur de passage 305

§41. Certains procédés d'évaluation des normes des puissances d'opérateurs .. 315
1. Conditions spectrales nécessaires pour que $\| R_h^p \|$ soient bornés (316). 2. Critère spectral de puissances bornées d'un opérateur auto-adjoint (317). 3. Critères de R_h auto-adjoint (318). 4. Evaluation des valeurs propres de l'opérateur R_h (319). 5. Choix du produit scalaire (320). 6. Critères de stabilité de Samarski (321).
Exercice 323

Chapitre 13. CRITÈRE SPECTRAL DE STABILITÉ DES PROBLÈMES AUX LIMITES ÉVOLUTIFS NON AUTO-ADJOINTS 324

§42. Spectre d'une famille d'opérateurs $\{R_h\}$ 324
1. Nécessité de perfectionner le critère spectral de stabilité (324). 2. Spectre d'une famille d'opérateurs (définition) (326). 3. Condition nécessaire de stabilité (326). 4. Notion de spectre de la famille d'opérateurs $\{R_h\}$ (discussion) (327). 5. Le critère nécessaire de stabilité est proche d'un critère suffisant (329).

§43. Algorithme de calcul du spectre d'une famille d'opérateurs aux différences pour les fonctions définies sur un segment maillé 331
1. Exemple caractéristique (331). 2. Algorithme de calcul du spectre dans le cas général (337).
Exercices 338

§44. Noyaux des spectres des familles d'opérateurs 338

§45. Sur la stabilité des algorithmes itératifs de résolution d'équations aux différences non auto-adjointes 341

Annexe. MÉTHODE DES CONDITIONS AUX LIMITES INTRINSÈQUES 344
1. Classe de systèmes d'équations aux différences (344). 2. Solution fondamentale (345). 3. Frontière d'un domaine discrétisé (345). 4. Analogues aux différences des formules intégrales de Cauchy et du type de Cauchy (346). 5. Conditions aux limites intrinsèques (347). 6. Opérateur projection frontière (348). 7. Problème aux limites général (348). 8. Idée à la base de la méthode des conditions aux limites intrinsèques (348). 9. Stabilité des conditions aux limites intrinsèques (348). 10. Une idée supplémentaire (349). 11. Comparaison de la méthode des conditions aux limites intrinsèques avec la méthode des équations intégrales singulières (351).

Commentaires 353

Bibliographie 358

Index 359

PRÉFACE

De nombreux problèmes des sciences de la nature se réduisent à des problèmes aux limites pour les équations différentielles qui pour être résolus sur ordinateurs doivent être remplacés approximativement par des schémas aux différences.

Le présent ouvrage se donne pour but d'initier à la théorie des schémas aux différences et est destiné aux étudiants des grandes Ecoles techniques, de l'Institut physico-technique et de l'Institut des ingénieurs physiciens de Moscou ainsi qu'à ceux des facultés de physique et de mathématiques. Certains chapitres présenteront probablement un intérêt pour les spécialistes du calcul numérique.

La structure de l'ouvrage a été conçue pour répondre aux exigences de ce public varié.

Le livre comprend cinq parties et un petit Annexe. Plusieurs parties (deux ou plus) constituent une propédeutique. La quantité de matériaux à étudier peut également être augmentée ou diminuée aux dépens des alinéas en petits caractères et des Exercices.

La bibliographie cite des livres et articles permettant d'approfondir de nombreuses questions de la théorie et des applications des schémas aux différences et de poursuivre les recherches bibliographiques.

L'ouvrage [5] constitue une introduction encore plus concise à la théorie des schémas aux différences.

On rencontre des références à des travaux originaux dans les cas peu fréquents où les résultats complémentaires sont donnés sans démonstration.

Les machines à calculer modernes et l'expérience acquise permettent de résoudre approximativement, grâce aux schémas aux différences, des problèmes très compliqués dont l'étude par d'autres méthodes s'avère plutôt difficile. On s'assure que la solution trouvée est exacte en appliquant le même schéma de calcul à quelques problèmes dont les solutions exactes sont connues d'avance, en comparant le résultat du calcul avec l'expérience physique dans le cadre des paramètres où celle-ci est possible et en recourant à d'autres procédés qui ne sont pas rigoureux sur le plan mathématique. Pour comprendre comment construire de bons schémas aux différences, l'on exa-

minera une série de problèmes modèles choisis de manière adéquate qui, tout en étant suffisamment simples pour qu'on puisse les analyser en détail avec toute la rigueur mathématique nécessaire, reflètent certains aspects du problème originel qu'on ne peut étudier rigoureusement soit qu'il soit trop complexe, soit que l'on manque de temps.

En insistant sur l'analyse rigoureuse des problèmes modèles nous avons également voulu inculquer au lecteur une idée juste de la relation entre la théorie et l'expérience sur machine lors de la formation de schémas aux différences pour des calculs pratiques.

Ce livre doit beaucoup au travail commun [4] des auteurs ainsi qu'aux conférences données par l'un d'eux au cours de plusieurs années à l'Institut physico-technique de Moscou. Ces conférences se sont inspirées de fréquentes discussions fécondes de leur auteur avec O. Biélotserkovski (qui en a été d'ailleurs le promoteur), avec V. Diatchenko, P. Fédorenko, O. Lokoutsievski, E. Schnol et L. Tchoudov.

Les auteurs expriment leurs remerciements les plus sincères à N. Bakhvalov et B. Rojdestvenski qui ont lu le manuscrit du présent livre et à tous ceux qui ont contribué à l'améliorer.

Les auteurs

INTRODUCTION

De nombreux problèmes théoriques et appliqués des sciences modernes de la nature conduisent à des équations différentielles, et ce n'est qu'après avoir résolu ces dernières qu'on peut estimer achevée l'étude d'un problème.

Dans certains cas on dispose d'une formule donnant la solution par l'intermédiaire de fonctions élémentaires bien étudiées. Mais la chose est impossible en principe. Aussi l'obtention de formules explicites ne saurait-elle être considérée comme un processus régulier aboutissant à la résolution d'équations différentielles. Il serait injuste de dire que cette approche analytique est complètement caduque. Elle reste un instrument nécessaire et fort puissant d'étude des problèmes simplifiés, dits modèles. En étudiant un problème modèle dûment choisi on peut faire des conclusions sur le comportement de la solution du problème initial.

A côté de cette façon d'agir analytique on utilise de plus en plus volontiers diverses méthodes de résolution numérique d'équations différentielles. Leur implantation à une vaste échelle a suivi l'apparition des machines à calculer rapides capables d'emmagasiner de grandes tables de nombres et d'effectuer sur ces derniers les opérations arithmétiques selon un programme donné. Etant donné ces facultés des machines, toute méthode numérique consiste à passer de la solution cherchée à une table inconnue de nombres et à indiquer la succession des opérations arithmétiques à effectuer. On peut par exemple chercher plusieurs premiers coefficients du développement de la solution en série entière ou trigonométrique. Le présent livre expose la théorie de la résolution numérique d'équations différentielles par la méthode des différences finies. L'idée de cette méthode numérique la plus universelle est de prendre pour l'ensemble cherché de nombres la table de valeurs de la solution aux points d'un ensemble appelé *réseau*. Le calcul de la table cherchée se fait à l'aide d'équations algébriques qui remplacent approximativement l'équation différentielle.

Illustrons-le sur l'exemple élémentaire d'un schéma aux différences pour le calcul approché de la solution de l'équation

$$u'(x)+Au(x) = 0$$

vérifiant la condition initiale $u(0) = 1$. Admettons que $h > 0$ et cherchons, au lieu de $u(x)$, la table de ses valeurs

$$u(0),\ u(h),\ u(2h),\ \ldots,\ u(nh),\ \ldots$$

Substituons à la dérivée le rapport aux différences approximant

$$\frac{u(x+h)-u(x)}{h}.$$

Le pas h de la table sera choisi suffisammant petit. Après cette substitution l'équation différentielle est remplacée par l'équation aux différences

$$\frac{u(x+h)-u(x)}{h}+Au(x) = 0$$

qui l'approche et qui permet de calculer approximativement la table cherchée. Pour cela mettons l'équation au différences sous forme de la formule de récurrence

$$u(x+h) = (1-Ah)\,u(x).$$

Posons successivement $x = 0, h, 2h, \ldots$, il vient

$$\begin{aligned} u(h) &= (1-Ah), \\ u(2h) &= (1-Ah)^2, \\ &\cdots\cdots\cdots \\ u(Nh) &= (1-Ah)^N, \\ &\cdots\cdots\cdots \end{aligned}$$

En choisissant $h = 1/N$, nous obtenons

$$u(1) = \left(1-\frac{A}{N}\right)^N$$

au lieu de la solution exacte

$$u(1) = e^{-A}.$$

Le cours d'analyse mathématique nous apprend cependant que pour un h suffisamment faible ou, ce qui revient au même, pour un N suffisamment grand, la quantité $(1-A/N)^N$ diffère peu de e^{-A}. Cela signifie que la solution approchée obtenue par ce schéma aux différences et dépendant du pas h converge avec la diminution de h vers la solution exacte de l'équation différentielle.

La même équation différentielle

$$u'(x)+Au(x) = 0$$

peut être approchée par une autre équation aux différences, à savoir par

$$\frac{u(x+h)-u(x-h)}{2h}+Au(x) = 0,$$

qui s'obtient en remplaçant la dérivée par le rapport aux différences

$$\frac{u(x+h)-u(x-h)}{2h}.$$

On construit pour l'équation différentielle

$$u''(x)+Au'(x)+Bu(x)=f(x)$$

un analogue aux différences en substituant à $u''(x)$ par exemple l'expression approchée suivante :

$$\frac{\dfrac{u(x+h)-u(x)}{h}-\dfrac{u(x)-u(x-h)}{h}}{h}=\frac{u(x+h)-2u(x)+u(x-h)}{h^2}.$$

Quant à la dérivée première $u'(x)$, elle peut être remplacée par une expression aux différences déjà utilisée. Il vient alors

$$\frac{u(x+h)-2u(x)+u(x-h)}{h^2}+A\,\frac{u(x+h)-u(x-h)}{2h}+Bu(x)=f(x).$$

S'il s'agit d'équations différentielles à coefficients variables, la construction de schémas aux différences ne se complique aucunement. On demande par exemple de calculer la solution de l'équation

$$u'(x)+A(x)\,u(x)=0,$$

où le coefficient A est fonction de x. On le fait à l'aide de l'équation aux différences

$$\frac{u(x+h)-u(x)}{h}+A(x)\,u(x)=0.$$

Les schémas aux différences « se jouent » de même des équations non linéaires. Ainsi, l'équation

$$u'(x)+\sin\left(xu(x)\right)=0$$

est susceptible d'être résolue approximativement moyennant le schéma

$$\frac{u(x+h)-u(x)}{h}+\sin\left(xu(x)\right)=0.$$

A considérer les exemples ci-dessus on pourrait penser que la formation d'un schéma aux différences et la résolution, ce schéma aidant, de l'équation différentielle ne présentent aucune difficulté. Il n'en est rien.

Dès les cas les plus simples, même en résolvant des équations linéaires à coefficients constants, il arrive souvent que la solution d'un schéma aux différences « raisonnable » ne converge pas, lors du passage à un réseau plus fin, vers la solution cherchée de l'équation différentielle. On conçoit qu'avec ce schéma on ne saurait calculer la fonction cherchée avec une précision aussi grande qu'on le veut.

Une fois le schéma convergent construit, il est nécessaire de calculer la solution du nouveau système d'équations algébriques par rapport à un grand nombre de valeurs inconnues de la fonction aux nœuds du réseau. Ce calcul s'avère ardu dans nombre de cas importants. On pallie parfois cet inconvénient en choisissant un schéma convergent de structure différente de sorte que le système d'équations linéaires qui surgit se résolve facilement de

façon exacte ; dans d'autres cas, on dispose des procédés de calcul approché des solutions des problèmes aux différences avec une précision donnée à l'avance.

Toute personne s'occupant de la résolution numérique des équations différentielles doit être au courant des difficultés soulevées par la construction et la mise en œuvre de schémas aux différences et des moyens d'en triompher.

PREMIÈRE PARTIE

ÉQUATIONS AUX DIFFÉRENCES ORDINAIRES

CHAPITRE PREMIER

ÉQUATIONS AUX DIFFÉRENCES DU PREMIER ET DU SECOND ORDRE. EXEMPLES DE SCHÉMAS AUX DIFFÉRENCES

§1. Equations aux différences simples

1. Equations aux différences. Dans l'introduction, nous avons construit pour l'équation différentielle du premier ordre

$$u'(x)+Au(x)=f(x)$$

deux schémas aux différences

$$\frac{u(x+h)-u(x)}{h}+Au(x)=f(x),$$

$$\frac{u(x+h)-u(x-h)}{2h}+Au(x)=f(x),$$

qui peuvent s'écrire respectivement

$$-\frac{1-Ah}{h}u(x)+\frac{1}{h}u(x+h)=f(x), \tag{1}$$

$$-\frac{1}{2h}u(x-h)+Au(x)+\frac{1}{2h}u(x+h)=f(x), \tag{2}$$

et pour l'équation différentielle du second ordre

$$u''(x)+Au'(x)+Bu(x)=f(x)$$

l'équation aux différences

$$\frac{u(x+h)-2u(x)+u(x-h)}{h^2}+A\,\frac{u(x+h)-u(x-h)}{2h}+Bu(x)=f(x),$$

ou, en modifiant l'écriture,

$$\frac{1}{h^2}\left(1-\frac{Ah}{2}\right)u(x-h)-\frac{1}{h^2}(2-Bh^2)\,u(x)+\frac{1}{h^2}\left(1+\frac{Ah}{2}\right)u(x+h)=f(x). \tag{3}$$

Les équations aux différences ci-dessus approchant les équations diffé-

rentielles les plus simples sont de l'un des deux types suivants :

$$au(x)+bu(x+h)=f(x), \tag{1'}$$

$$au(x-h)+bu(x)+cu(x+h)=f(x). \tag{2'}$$

Si la suite de points partageant l'axe Ox en des segments de longueur h est numérotée de gauche à droite de façon que $x_n = x_{n-1}+h$ et si on désigne $u(x_n)$ par u_n et $f(x_n)$ par f_n, nos schémas aux différences se mettent sous forme d'équations

$$au_n+bu_{n+1}=f_n, \tag{4}$$

$$au_{n-1}+bu_n+cu_{n+1}=f_n. \tag{5}$$

Aux §§1-4 nous étudierons des équations aux différences de la forme (4) et (5) sans préciser si elles constituent les schémas aux différences pour des équations différentielles.

Les inconnues u_n des équations (4) et (5) forment une suite $\{u_n\}$:

$$\ldots, u_{-3}, u_{-2}, u_{-1}, u_0, u_1, u_2, u_3, \ldots$$

Dans l'exposé suivant nous comparerons souvent cette suite avec une suite de points numérotés par les nombres

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$$

ou, comme on dit encore, avec un réseau.

La suite $\{u_n\}$ peut être assimilée à une fonction u définie aux points d'un réseau, et u_k est la valeur de u au point de numéro k. La fig. 1 est le graphe

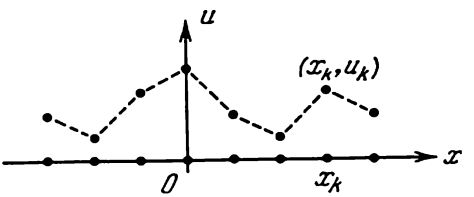


Fig. 1

d'une fonction u définie sur un réseau. Il s'agit de l'ensemble des points (x_k, u_k) dans le plan Oxu.

Du moment que nous ne nous intéressons pas au lien entre les équations aux différences et les équations différentielles, la distance entre deux points voisins ne doit pas être forcément h. Cette distance peut être arbitraire, par exemple égale à l'unité, et on peut prendre pour x_0 le point 0. La fonction u sera alors définie aux points de coordonnées nombres entiers $x_k = k$.

Supposons, pour simplifier, que les coefficients a, b, c des équations (4), (5) sont constants. En disant que les équations considérées sont à coefficients constants, nous avons en vue l'indépendance de ces coefficients par rapport

au numéro n ; ainsi, l'équation

$$u_{n-1}+5\sqrt{n}u_n+u_{n+1} = 0$$

n'est pas à coefficients constants.

Partout dans la suite les coefficients a, b de l'équation (4) et a, c de l'équation (5) seront supposés différents de zéro.

La suite $\{f_n\}$ s'appelle *second membre* des équations considérées.

La suite $\{u_n\}$ étant assujettie à la seule condition d'être définie en tous les points entiers n, $-\infty < n < \infty$, on voit sans peine que les équations (4) et (5) admettent nombre de solutions. Ainsi, $qu_n-u_{n+1} = 0$ a pour solutions $u_n \equiv 0$ et $u_n = q^n$.

Pour pouvoir mettre en évidence *la* solution de l'équation (4)

$$au_n+bu_{n+1} = f_n,$$

il suffit de définir sa valeur en un seul point entier m, i.e. définir u_m. En effet, l'équation (4) peut s'écrire comme la formule de récurrence

$$u_{n+1} = \frac{1}{b}(f_n-au_n)$$

à partir de laquelle, pour $n = m, m+1, \ldots$, on définit de proche en proche $u_{m+1}, u_{m+2}, \ldots$, i.e. tous les u_n pour $n > m$. Mettant notre équation sous forme d'une autre formule de récurrence :

$$u_{n-1} = \frac{1}{a}(f_n-bu_n),$$

nous pouvons de même définir tous les u_n pour $n < m$.

Pour mettre en évidence *la* solution de l'équation (5)

$$au_{n-1}+bu_n+cu_{n+1} = f_n$$

il suffit de se donner d'une façon arbitraire les valeurs de u en deux points entiers successifs quelconques, de se donner par exemple u_{m-1} et u_m. L'équation considérée pouvant être récrite sous forme de deux formules de récurrence suivantes :

$$u_{n+1} = \frac{1}{c}(f_n-bu_n-au_{n-1}),$$

$$u_{n-1} = \frac{1}{a}(f_n-bu_n-cu_{n+1}),$$

la démonstration en est immédiate.

2. Ordre d'une équation aux différences. Répétons le résultat que nous venons d'obtenir et formulons la notion d'ordre pour les équations aux différences (4) et (5).

Pour mettre en évidence une solution unique de l'équation (4)

$$au_n+bu_{n+1} = f_n,$$

il suffit de définir la valeur de u en un seul point. Une telle équation s'appelle

équation du premier ordre. Pour trouver l'unique solution de l'équation (5)

$$au_{n-1}+bu_n+cu_{n+1}=f_n,$$

il suffit de définir les valeurs de la solution en deux points successifs et une telle équation s'appelle donc *équation du second ordre*.

On pourrait de même considérer l'équation très simple

$$au_n=f_n,\ a\neq 0,$$

dont la solution est définie de façon unique sans que la suite $\{u_n\}$ soit assujettie à des conditions préliminaires. Il n'est que naturel de parler dans ce cas d'une *équation d'ordre zéro*.

Le schéma aux différences simple (1) pour l'équation différentielle du premier ordre $u'+Au=f$ est une équation aux différences du premier ordre.

Le schéma (3) pour l'équation différentielle du second ordre $u''+Au'+$ $+Bu=f$ est d'ordre 2.

En examinant le schéma (2)

$$-\frac{1}{2h}u(x-h)+Au(x)+\frac{1}{2h}u(x+h)=f(x)$$

pour l'équation $u'+Au=f$, on constate que l'ordre de l'équation aux différences peut être *supérieur* à celui de l'équation différentielle (nous avons en l'occurrence une équation différentielle d'ordre 1, et l'équation aux différences correspondante est d'ordre 2).

3. Solution générale d'une équation aux différences. Etudions maintenant la structure des solutions des équations aux différences considérées. Prenons pour commencer l'équation homogène

$$a\bar{u}_n+b\bar{u}_{n+1}=0 \tag{6}$$

et notons Y_n la solution de l'équation (6) vérifiant la condition initiale $Y_0=1$. Visiblement, $\bar{u}_n=\alpha Y_n$ est également solution de l'équation homogène pour une constante arbitraire α. On montre aisément que toute solution de l'équation homogène (6) s'obtiendra sous cette forme. En effet, chaque solution se définit de façon unique par sa valeur pour $n=0$. Or, la solution $\bar{u}_n$ prenant une valeur donnée $\bar{u}_0$ s'obtient selon la formule $\bar{u}_n=\alpha Y_n$ si α y est remplacé par $\bar{u}_0$.

Passons à l'équation non homogène (4)

$$au_n+bu_{n+1}=f_n.$$

Soient $\{\tilde{u}_n\}$ et $\{u_n^*\}$ ses deux solutions quelconques. Retranchons terme à terme les égalités

$$a\tilde{u}_n+b\tilde{u}_{n+1}=f_n,$$
$$au_n^*+bu_{n+1}^*=f_n;$$

nous voyons que la différence $\tilde{u}_n-u_n^*=\bar{u}_n$ vérifie l'équation homogène (6) $a\bar{u}_n+b\bar{u}_{n+1}=0$. Aussi toute solution $\{\tilde{u}_n\}$ peut-elle prendre la forme

$$\tilde{u}_n=u_n^*+\bar{u}_n=u_n^*+\alpha Y_n$$

pour un choix convenable de la constante α. D'autre part, on vérifie aisément que, pour α quelconque, la formule $u_n = u_n^* + \alpha Y_n$ définit une solution de l'équation non homogène

$$\begin{aligned} au_n + bu_{n+1} &= a(u_n^* + \alpha Y_n) + b(u_{n+1}^* + \alpha Y_{n+1}) = \\ &= (au_n^* + bu_{n+1}^*) + \alpha(aY_n + bY_{n+1}) = f_n + \alpha \cdot 0 = f_n. \end{aligned}$$

Nous avons donc montré que la solution générale de l'équation homogène (6)

$$a\bar{u}_n + b\bar{u}_{n+1} = 0$$

est de la forme

$$\bar{u}_n = \alpha Y_n,$$

où Y_n est solution particulière de cette équation, qui vérifie la condition initiale $Y_0 = 1$, et α une constante arbitraire. La solution générale de l'équation non homogène (4)

$$au_n + bu_{n+1} = f_n$$

s'obtient comme

$$u_n = u_n^* + \alpha Y_n,$$

u_n^* étant une solution particulière de (4) et α une constante arbitraire.

En raisonnant de même sur une équation aux différences du second ordre, on aboutit à une assertion analogue. Nous nous abstiendrons de ces raisonnements (nous en laissons le soin au lecteur) en nous bornant au résultat :

La solution générale de l'équation aux différences homogène

$$a\bar{u}_{n-1} + b\bar{u}_n + c\bar{u}_{n+1} = 0 \tag{7}$$

peut être cherchée sous la forme

$$\bar{u}_{n-1} = \alpha Y_n + \beta Z_n,$$

où Y_n et Z_n sont deux solutions particulières de (7) vérifiant les conditions initiales

$$\begin{aligned} Y_0 &= 1, \quad Y_1 = 0, \\ Z_0 &= 0, \quad Z_1 = 1, \end{aligned}$$

et α et β des constantes quelconques.

La solution générale de l'équation non homogène (5)

$$au_{n-1} + bu_n + cu_{n+1} = f_n$$

est susceptible de prendre la forme

$$u_n = u_n^* + \alpha Y_n + \beta Z_n,$$

u_n^* étant une solution particulière de cette équation non homogène.

Tous les résultats et raisonnements du présent paragraphe s'étendent

sans aucune modification aux équations aux différences à coefficients variables. Nous préférons cependant passer outre afin de ne pas alourdir l'exposé par des détails sans importance.

Exercices

1. Démontrer que la solution générale de l'équation aux différences homogène

$$a_n u_n + b_n u_{n+1} = 0$$

à coefficients variables $a_n \neq 0$, $b_n \neq 0$ peut s'écrire $u_n = \alpha y_n$, où y_n est une solution particulière arbitraire qui s'annule pour certains n, et α une constante.

2. Démontrer que la solution générale de l'équation aux différences homogène d'ordre 2

$$a_n u_{n-1} + b_n u_n + c_n u_{n+1} = 0$$

à coefficients variables $a_n \neq 0$, $c_n \neq 0$ peut être recherchée sous la forme

$$u_n = \alpha y_n + \beta z_n,$$

y_n et z_n étant deux solutions particulières quelconques de cette équation telles que le déterminant

$$\begin{vmatrix} y_0 & y_1 \\ z_0 & z_1 \end{vmatrix}$$

soit non nul.

3. Soient y_n et z_n deux solutions particulières quelconques de l'équation aux différences d'ordre 2 de l'Exercice 2. Démontrer que le déterminant

$$\begin{vmatrix} y_n & y_{n+1} \\ z_n & z_{n+1} \end{vmatrix} = y_n z_{n+1} - z_n y_{n+1}$$

est soit nul pour tout n, soit non nul pour tous les n.

4. En combien de points successifs doit-on définir les valeurs de la solution de l'équation aux différences

$$a u_n + b u_{n+1} + c u_{n+2} + d u_{n+3} = f_n,$$

$a_n \neq 0$, $d \neq 0$, pour qu'il existe une solution $\{u_n\}$ et une seule prenant en ces points des valeurs données ? Quel est l'ordre de l'équation aux différences proposée ?

§2. Equation aux différences du premier ordre

Nous allons obtenir la formule donnant la solution générale de l'équation aux différences du premier ordre à coefficients constants

$$a u_n + b u_{n+1} = f_n,$$

les contraintes imposées à f_n étant assez faibles.

La solution générale peut être (voir §1) de la forme

$$u_n = u_n^* + \alpha Y_n = u_n^* + \alpha \left(-\frac{a}{b}\right)^n,$$

u_n^* étant une solution particulière et α une constante arbitraire.

La recherche de la solution générale s'est donc ramenée à celle d'une solution particulière u_n^*.

1. Solution fondamentale. Construisons d'abord la solution pour un second membre donné de façon spéciale

$$f_n = \begin{cases} 0, & n \neq 0, \\ 1, & n = 0. \end{cases}$$

Cette fonction est d'ordinaire désignée par le symbole de Kronecker

$$\delta_k^n = \begin{cases} 0, & n \neq k, \\ 1, & n = k. \end{cases}$$

Alors $f_n = \delta_0^n$.

La solution de l'équation

$$au_n + bu_{n+1} = \delta_0^n$$

sera notée G_n :

$$aG_n + bG_{n+1} = \delta_0^n. \tag{1}$$

La solution G_n s'appelle *solution fondamentale* de l'équation

$$au_n + bu_{n+1} = f_n,$$

vu qu'elle sert à exprimer ses solutions particulières pour des seconds membres f_n différents, assez arbitraires (voir p. 24).

On demande de trouver une solution des équations des trois groupes suivants :

I. $aG_n + bG_{n+1} = 0$ pour $n \leqslant -1$.

II. $aG_0 + bG_1 = 1$.

III. $aG_n + bG_{n+1} = 0$ pour $n \geqslant 1$.

Soit $G_n = 0$ lorsque $n \leqslant 0$. Alors toutes les équations du groupe I sont vérifiées. Moyennant l'équation II on trouve $G_1 = 1/b$. Les équations du groupe III peuvent s'écrire comme la formule récurrentielle $G_{n+1} = -\frac{a}{b} G_n$ qui permet de trouver successivement

$$G_2 = \frac{1}{b}\left(-\frac{a}{b}\right) = -\frac{1}{a}\left(-\frac{a}{b}\right)^2,$$

$$G_3 = \frac{1}{b}\left(-\frac{a}{b}\right)^2 = -\frac{1}{a}\left(-\frac{a}{b}\right)^3,$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$G_n = -\frac{1}{a}\left(-\frac{a}{b}\right)^n \quad \text{pour } n \geqslant 1.$$

Résumons les formules donnant G_n :

$$G_n = \begin{cases} 0 & \text{pour } n \leqslant 0, \\ -\frac{1}{a}\left(-\frac{a}{b}\right)^n & \text{pour } n \geqslant 1. \end{cases} \tag{2}$$

C'est une des solutions de (1). Faisons la somme de cette solution et de la solution générale $A\left(-\frac{a}{b}\right)^n$ de l'équation homogène associée $au_n+bu_{n+1}=0$, nous obtenons la solution générale de l'équation (1) :

$$G_n = \begin{cases} A\left(-\frac{a}{b}\right)^n & \text{pour } n \leqslant 0, \\ \left(A-\frac{1}{a}\right)\left(-\frac{a}{b}\right)^n & \text{pour } n \geqslant 1. \end{cases} \tag{3}$$

La solution fondamentale (2) s'obtient à partir de la formule générale (3) pour $A = 0$.

2. Condition de borne de la solution fondamentale. Si $|a/b| = 1$ et quelle que soit la constante A, la solution fondamentale G_n obtenue est bornée en valeur absolue tant pour $n \to +\infty$ pue pour $n \to -\infty$. Dégageons de la formule générale (3) la solution fondamentale bornée G_n pour $|a/b| \neq 1$. Si $|a/b| < 1$, alors $|-a/b|^n$ croît indéfiniment lorsque n tend vers l'infini négatif. C'est pourquoi, on n'obtient la solution bornée que lorsque $A = 0$ (fig. 2, *a*). Cette solution est définie par la formule (2).

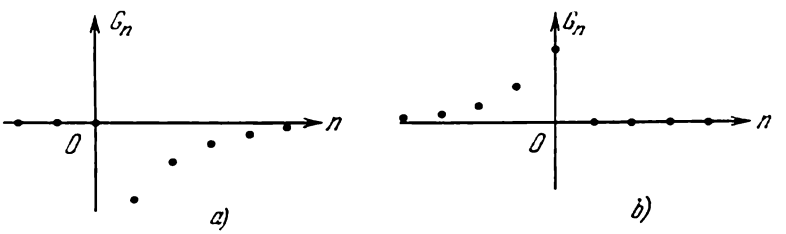


Fig. 2

Si $|a/b| > 1$, la solution bornée n'a lieu que lorsque $A = 1/a$ (fig. 2, *b*) :

$$G_n = \begin{cases} \frac{1}{a}\left(-\frac{a}{b}\right)^n, & n \leqslant 0, \\ 0, & n \geqslant 1. \end{cases} \tag{4}$$

3. Solution particulière. Une solution particulière de l'équation

$$au_n+bu_{n+1}=f_n, \tag{5}$$

avec f_n arbitraire, peut être cherchée sous forme de série

$$u_n = \sum_{k=-\infty}^{\infty} G_{n-k}f_k \tag{6}$$

(G_n étant une solution fondamentale) à condition que cette série converge.

Pour le montrer utilisons l'égalité

$$aG_{n-k}+bG_{n-k+1} = \delta_0^{n-k}(=\delta_k^n)$$

qui s'obtient de (1) en y changeant partout n en $n-k$. Portons la série convergente (6) dans le premier membre de (5), il vient

$$au_n+bu_{n+1} = a\sum_{k=-\infty}^{\infty} G_{n-k}f_k+b\sum_{k=-\infty}^{\infty} G_{n-k+1}f_k =$$

$$= \sum_{k=-\infty}^{\infty} (aG_{n-k}+bG_{n-k+1})f_k = \sum_{k=-\infty}^{\infty} \delta_k^n f_k = f_n.$$

En l'absence de toute hypothèse sur le comportement du second membre f_k de l'équation aux différences, la série (6) peut s'avérer divergente. En effet, si on pose $f_k = (-a/b)^k$, alors

$$G_{n-k}f_k = \begin{cases} A\left(-\dfrac{a}{b}\right)^n & \text{pour } n \leqslant k, \\ \left(A-\dfrac{1}{a}\right)\left(-\dfrac{a}{b}\right)^n & \text{pour } n \geqslant k+1, \end{cases}$$

et la série (6) contient, pour n fixe, une infinité de termes identiques non nuls.

THÉORÈME. *Soient* $|a/b| \neq 1$, G_n *la solution fondamentale bornée et* f_k *bornés en module, i.e.* $|f_k| < F$. *Alors la série*

$$u_n = \sum_{k=-\infty}^{\infty} G_{n-k}f_k$$

est manifestement convergente.

DÉMONSTRATION. La preuve ne sera faite que pour $|a/b| > 1$. Le cas contraire ne présentera alors pas de difficultés pour le lecteur.

Dans nos hypothèses, chaque terme de la série

$$u_n = \sum_{k=-\infty}^{\infty} G_{n-k}f_k = \sum_{k=n}^{\infty} \left[\frac{1}{a}\left(-\frac{a}{b}\right)^{n-k}\right]f_k$$

peut être majoré en valeur absolue par le terme d'une progression géométrique convergente

$$\left|-\frac{1}{a}\left(-\frac{a}{b}\right)^{n-k}f_k\right| \leqslant \frac{F}{|a|}\left|\frac{a}{b}\right|^{n-k}.$$

D'où la convergence de la série (6) et l'estimation

$$|u_n| \leqslant \frac{F}{|a|}\sum_{k=n}^{\infty}\left|\frac{b}{a}\right|^{k-n} = \frac{F}{|a|-|b|} \tag{7}$$

montrant que la solution (6) est bornée.

C'est la seule solution bornée de l'équation

$$au_n+bu_{n+1} = f_n$$

puisque toute solution s'obtient de (6) par adjonction d'une solution $\bar{u}_n =$

$= \alpha\left(-\frac{a}{b}\right)^n$ de l'équation homogène associée. La solution $\{\bar{u}\}$ est bornée en tant que différence de deux solutions bornées, ce qui n'est possible que pour $\alpha = 0$.

Exercices

1. Trouver la solution générale de l'équation

$$2u_n - u_{n+1} = 5^n.$$

SOLUTION. La solution générale de l'équation homogène correspondante $2\bar{u}_n - \bar{u}_{n+1} = 0$ est de la forme $\bar{u}_n = \alpha(2)^n$. Cherchons une solution particulière u_n^* de la forme $u_n^* = C5^n$ à coefficient indéterminé. En portant $u_n^* = C5^n$ dans l'équation on obtient

$$(2\cdot 5^n - 5^{n+1})C = 5^n; \quad C = -1/3.$$

Donc

$$u_n = -\frac{5^n}{3} + \alpha(2)^n.$$

(Notons que u_n^* ne peut s'obtenir comme la série (6) puisque, son terme général ne tendant pas vers zéro, la série diverge.)

2. Trouver une solution particulière u_n^* de l'équation

$$2u_n - u_{n+1} = (2)^n.$$

INDICATION. Chercher la solution sous la forme $u_n^* = Cn\cdot(2)^n$.

3. Trouver les solutions particulières u_n^* de l'équation

$$2u_n - u_{n+1} = f_n$$

pour le second membre de la forme

$$\text{a) } f_n \equiv 1, \quad \text{b) } f_n = n, \quad \text{c) } f_n = n^2, \quad \text{d) } f_n = 1+2n-n^2.$$

4. Trouver les solutions particulières u_n^* de l'équation

$$u_n - u_{n+1} = f_n$$

dont le second membre a la forme spéciale suivante:

$$\text{a) } f_n = 1, \quad \text{b) } f_n = n, \quad \text{c) } f_n = n^2.$$

§3. Equation aux différences du second ordre

Nous allons obtenir la formule donnant la solution générale de l'équation aux différences non homogène à coefficients constants

$$au_{n-1} + bu_n + cu_{n+1} = f_n. \tag{1}$$

Nous avons établi (§1) que sa solution générale s'écrit

$$u_n = u_n^* + \bar{u}_n, \tag{2}$$

u_n^* étant une solution particulière de l'équation non homogène donnée et

$$\bar{u}_n = \alpha Y_n + \beta Z_n$$

la solution générale de l'équation homogène correspondante

$$au_{n-1} + bu_n + cu_{n+1} = 0. \tag{3}$$

Trouvons d'abord la formule de la solution générale de l'équation homogène (3), puis la solution fondamentale et une solution particulière de l'équation non homogène.

1. Solution générale de l'équation homogène. En nous rappelant que, dans le cas de l'équation aux différences du premier ordre, nous avons eu une solution particulière de la forme $u_n = q^n$, essayons, cette fois encore, de chercher la solution particulière sous forme d'une progression géométrique. Portons $u_n = q^n$ dans l'équation aux différences et assurons-nous que c'est effectivement une solution si q est racine de l'équation du deuxième degré

$$a+bq+cq^2 = 0 \tag{4}$$

dite *équation caractéristique*. Cette dernière équation peut avoir ses racines distinctes ou non. Considérons les deux cas. Si les racines q_1 et q_2 de l'équation caractéristique en question sont distinctes, il est possible de trouver même deux solutions particulières indépendantes

$$u_n^{(1)} = q_1^n, \quad u_n^{(2)} = q_2^n$$

sous forme d'une progression géométrique. La combinaison linéaire

$$\bar{u}_n = \alpha u_n^{(1)}+\beta u_n^{(2)} = \alpha q_1^n+\beta q_2^n \tag{5}$$

à coefficients constants arbitraires α et β de ces deux solutions est encore solution de l'équation homogène. Montrons qu'il s'agit de la solution générale.

En effet, toute solution particulière $\bar{u}_n$ de l'équation homogène qui prend pour $n = 0$ et $n = 1$ des valeurs $\bar{u}_0$ et $\bar{u}_1$ quelconques données à l'avance peut s'obtenir sous la forme (5). Il suffit de définir α et β à partir des égalités

$$\alpha+\beta = \bar{u}_0,$$
$$\alpha q_1+\beta q_2 = \bar{u}_1,$$

i.e. de poser

$$\alpha = \frac{\bar{u}_0 q_2-\bar{u}_1}{q_2-q_1}, \quad \beta = \frac{\bar{u}_1-\bar{u}_0 q_1}{q_2-q_1}.$$

En particulier, Y_n et Z_n définis au §1 comme solutions de l'équation homogène telles que

$$Y_0 = 1, \quad Y_1 = 0,$$
$$Z_0 = 0, \quad Z_1 = 1$$

sont de la forme

$$\left.\begin{aligned} Y_n &= \frac{q_2}{q_2-q_1} q_1^n - \frac{q_1}{q_2-q_1} q_2^n, \\ Z_n &= -\frac{1}{q_2-q_1} q_1^n + \frac{1}{q_2-q_1} q_2^n. \end{aligned}\right\} \tag{6}$$

A l'examen des formules (6) on constate qu'elles tombent en défaut lorsque $q_1 = q_2$ (racine multiple).

Si $q_1 = q_2$, une solution particulière peut s'écrire toujours $u_n = q_1^n$. On trouve une deuxième en faisant $u_n = y_n q_1^n$ dans (3), ce qui donne pour y_n

$$ay_{n-1}+bq_1y_n+cq_1^2y_{n+1} = 0.$$

On sait que a/c est égal au produit et b/c à la somme de signe contraire des racines de l'équation caractéristique (4). Puisque les deux racines valent q_1, on a

$$\frac{a}{c} = q_1^2, \quad \frac{b}{c} = -2q_1,$$

et l'équation aux différences pour y_n se récrit

$$cq_1^2y_{n-1}-2cq_1^2y_n+cq_1^2y_{n+1} = 0$$

ou, plus brièvement,

$$y_{n-1}-2y_n+y_{n+1} = 0.$$

En mettant cette équation sous la forme

$$y_{n-1}-y_n = y_n-y_{n+1},$$

on voit que la différence $y_{n-1}-y_n$ ne change pas avec la variation de n. Ainsi, la solution est une progression arithmétique arbitraire. Etant donné qu'il nous suffit de trouver une seule solution, prenons la progression arithmétique $y_n = n$. Comme nous avons cherché u_n sous la forme $u_n = y_n q_1^n$, il existe parmi les solutions de l'équation $au_{n-1}+bu_n+cu_{n+1} = 0$ une solution

$$u_n^{(2)} = nq_1^n.$$

Ainsi, dans le cas des racines multiples $q_1 = q_2$, nous avons obtenu, en plus de la solution particulière $u_n^{(1)} = q_1^n$, une deuxième solution particulière indépendante $u_n^{(2)} = nq_1^n$.

La combinaison linéaire

$$\bar{u}_n = \alpha q_1^n+\beta nq_1^n$$

à coefficients constants arbitraires est encore solution de l'équation homogène et une solution particulière s'obtient de cette formule par un choix convenable de α et β. En particulier, dans le cas des racines multiples, les solutions particulières Y_n et Z_n ont la forme

$$\left.\begin{aligned} Y_n &= q_1^n-nq_1^n, \\ Z_n &= \frac{1}{q_1}nq_1^n = nq_1^{n-1}. \end{aligned}\right\} \tag{7}$$

Fait intéressant, les formules (7) peuvent s'obtenir à partir de (6), donnant Y_n et Z_n dans le cas de l'équation caractéristique ayant toutes ses racines simples. Nous avons eu pour Y_n et Z_n les égalités

$$Y_n = \frac{q_2}{q_2-q_1}q_1^n-\frac{q_1}{q_2-q_1}q_2^n = -q_1q_2\frac{q_2^{n-1}-q_1^{n-1}}{q_2-q_1},$$

$$Z_n = -\frac{1}{q_2-q_1}q_1^n+\frac{1}{q_2-q_1}q_2^n = \frac{q_2^n-q_1^n}{q_2-q_1}.$$

Faisons la racine q_2 tendre vers la racine q_1. Ceci étant, les expressions

$$\frac{q_2^{n-1}-q_1^{n-1}}{q_2-q_1} \quad \text{et} \quad \frac{q_2^n-q_1^n}{q_2-q_1}$$

tendent vers certaines limites, à savoir vers $(n-1)q_1^{n-2}$ et nq_1^{n-1} respectivement. Nous voyons donc que lorsqu'il s'agit des racines multiples, les solutions Y_n et Z_n prennent la forme (7).

Ainsi, nous avons construit Y_n et Z_n dans tous les cas possibles de a et c non nuls.

Nous avons montré par là même qu'on peut toujours écrire sous forme explicite toute solution de l'équation aux différences homogène du second ordre que nous sommes en train d'étudier.

Le cas où l'équation $a+bq+cq^2 = 0$ à coefficients a, b, c réels possède deux racines q_1, q_2 complexes conjuguées mérite qu'on s'y arrête. Montrons que la solution générale de l'équation homogène (3) s'écrit alors

$$\bar{u}_n = \gamma_1\left(\sqrt{\frac{a}{c}}\right)^n \cos n\varphi+\gamma_2\left(\sqrt{\frac{a}{c}}\right)^n \sin n\varphi, \tag{8}$$

avec φ défini par l'égalité

$$\cos\varphi = -\frac{b}{2\sqrt{ac}}$$

et γ_1 et γ_2 des constantes quelconques.

Explicitons q_1 et q_2:

$$q_1 = \sqrt{\frac{a}{c}}\left[-\frac{b}{2\sqrt{ac}}+i\sqrt{1-\left(\frac{b}{2\sqrt{ac}}\right)^2}\right],$$

$$q_2 = \sqrt{\frac{a}{c}}\left[-\frac{b}{2\sqrt{ac}}-i\sqrt{1-\left(\frac{b}{2\sqrt{ac}}\right)^2}\right].$$

Les racines étant complexes, on a $\frac{a}{c} > 0$, $\left|\frac{b}{2\sqrt{ac}}\right| < 1$. On peut donc noter

$$-\frac{b}{2\sqrt{ac}} = \cos\varphi, \quad \sqrt{1-\left(\frac{b}{2\sqrt{ac}}\right)^2} = \sin\varphi$$

et q_1 et q_2 s'écrivent

$$q_1 = \sqrt{\frac{a}{c}}(\cos\varphi+i\sin\varphi),$$

$$q_2 = \sqrt{\frac{a}{c}}(\cos\varphi-i\sin\varphi).$$

Portons ces valeurs de q_1 et q_2 dans la formule (5).

Lorsque $\alpha = \beta = 1/2$, nous obtenons une solution particulière

$$u_n^{(1)} = \left(\sqrt{\frac{a}{c}}\right)^n \cos n\varphi$$

et, lorsque $\alpha = 1/(2i)$, $\beta = -1/(2i)$,

$$u_n^{(2)} = \left(\sqrt{\frac{a}{c}}\right)^n \sin n\varphi.$$

La combinaison linéaire à coefficients γ_1 et γ_2 constants arbitraires de ces solutions particulières donne la solution générale (8). (Le lecteur vérifiera aisément qu'il est possible de mettre sous cette forme la solution particulière (8) prenant pour $n = 0$ et $n = 1$ des valeurs quelconques données à l'avance.)

2. Solution générale de l'équation non homogène. Solution fondamentale. Passons à l'étude de l'équation aux différences non homogène

$$au_{n-1}+bu_n+cu_{n+1} = f_n, \tag{9}$$

et bornons-nous à un cas important pour l'exposé suivant, cas où aucune racine de l'équation caractéristique (4) n'est égale en module à l'unité : $|q_1| \neq 1$, $|q_2| \neq 1$. Commençons par chercher la solution de l'équation non homogène (9) dont le second membre a une forme spéciale

$$f_n = \delta_0^n = \begin{cases} 0, & n \neq 0, \\ 1, & n = 0. \end{cases}$$

Cette solution sera notée G_n et appelée *solution fondamentale*. Nous voulons trouver une solution fondamentale *bornée*, i.e. une solution bornée des équations des trois groupes suivants :

I. $aG_{n-1}+bG_n+cG_{n+1} = 0$ pour $n \leqslant -1$.

II. $aG_{-1}+bG_0+cG_1 = 1$.

III. $aG_{n-1}+bG_n+cG_{n+1} = 0$ pour $n \geqslant 1$.

Examinons d'abord le cas des racines distinctes : $q_1 \neq q_2$.

La solution générale de l'équation homogène (3) est alors de la forme

$$u_n = \alpha q_1^n+\beta q_2^n.$$

C'est pourquoi, toute solution particulière G_n de l'équation homogène I s'écrit

$$G_n = \alpha' q_1^n+\beta' q_2^n \quad \text{pour } n \leqslant 0,$$

α' et β' étant des constantes dûment choisies. De même, une solution particulière G_n, $n \geqslant 0$, de l'équation homogène III s'obtient comme

$$G_n = \alpha'' q_1^n+\beta'' q_2^n \quad \text{pour } n \geqslant 0$$

avec des constantes α'' et β'' correspondantes.

Dans le cas étudié $q_1 \neq q_2$, $|q_1| \neq 1$, $|q_2| \neq 1$, quatre possibilités peuvent se présenter :

a) $|q_1| < 1$, $|q_2| > 1$;

b) $|q_1| < 1$, $|q_2| < 1$;

c) $|q_1| > 1, \quad |q_2| < 1;$

d) $|q_1| > 1, \quad |q_2| > 1.$

Construisons la solution fondamentale bornée G_n en nous plaçant dans le cas a). G_n étant borné pour $n \to -\infty$, on a $\alpha' = 0$, et, G_n l'étant pour $n \to +\infty$, $\beta'' = 0$. Aussi

$$G_n = \begin{cases} \beta' q_2^n & \text{pour } n \leqslant 0, \\ \alpha'' q_1^n & \text{pour } n \geqslant 0. \end{cases}$$

Lorsque $n = 0$, les deux dernières formules fournissent une même valeur G_0. D'où $\beta' = \alpha''$. Choisissons β' par la condition de validité de l'équation II :

$$a\beta' q_2^{-1} + b\beta' + c\beta' q_1 = 1, \quad \beta' = \frac{1}{aq_2^{-1} + b + cq_1}.$$

Le dénominateur de cette fraction est différent de zéro :

$$aq_2^{-1} + b + cq_1 = (aq_2^{-1} + b + cq_2) + c(q_1 - q_2) = c(q_1 - q_2) \neq 0.$$

Donc,

$$G_n = \begin{cases} \dfrac{1}{aq_2^{-1} + b + cq_1} q_2^n, & n \leqslant 0, \\ \dfrac{1}{aq_2^{-1} + b + cq_1} q_1^n, & n \geqslant 0. \end{cases}$$

Nous avons construit la solution fondamentale bornée demandée (fig. 3, *a*).

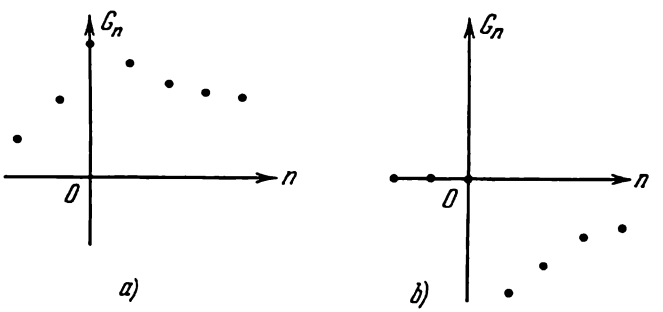


Fig. 3

Notons (ce fait va nous servir dans la suite) que dans les conditions

$$\left.\begin{aligned} &\max(|a|, |b|, |c|) \geqslant B > 0, \\ &|q_1| < 1 - \frac{\theta}{2}, \quad |q_2^{-1}| < 1 - \frac{\theta}{2}, \end{aligned}\right\} \tag{10}$$

avec $B > 0$ et $\theta > 0$ des nombres, on a l'estimation

$$|G_n| \leqslant \frac{4}{B\theta}\left(1 - \frac{\theta}{2}\right)^{|n|}. \tag{11}$$

En effet, en vertu de la première condition (10), on a forcément soit $|a| > B/4$, soit $|c| > B/4$, soit encore $\sqrt{b^2-4ac} \geqslant \sqrt{B^2-B^2/4} > B/2$.

On a également les relations évidentes

$$aq_2^{-1}+b+cq_1 = c(q_1-q_2) = a(q_2^{-1}-q_1^{-1}) = \sqrt{b^2-4ac},$$

$$|q_1-q_2| \geqslant |q_2|-|q_1| \geqslant \frac{1}{1-\theta/2}-(1-\theta/2) = \theta\,\frac{2-\theta/2}{2-\theta} > \theta,$$

$$|q_2^{-1}-q_1^{-1}| > \theta,$$

qui entraînent l'estimation

$$|aq_2^{-1}+b+cq_1| \geqslant \frac{B\theta}{4}$$

et l'inégalité (11).

Dans le cas b), la condition que G_n est borné pour $n \to -\infty$ entraîne $\alpha' = \beta' = 0$, de sorte que

$$G_n = \begin{cases} 0 & \text{pour } n \leqslant 0, \\ \alpha''q_1^n+\beta''q_2^n & \text{pour } n \geqslant 0. \end{cases}$$

La condition $G_0 = 0$ implique $\alpha'' = -\beta''$. Le coefficient α'' est choisi de façon que soit satisfaite l'équation II :

$$\alpha'' = -\frac{1}{c(q_1-q_2)}.$$

La solution fondamentale bornée (fig. 3, *b*) dans le cas b) est de la forme

$$G_n = \begin{cases} 0 & \text{pour } n \leqslant 0, \\ \dfrac{1}{c(q_1-q_2)}(q_1^n-q_2^n) & \text{pour } n \geqslant 0. \end{cases}$$

S'agissant du cas c) on a par analogie avec a) :

$$G_n = \begin{cases} \dfrac{1}{aq_1^{-1}+b+cq_2}\,q_1^n & \text{pour } n \leqslant 0, \\ \dfrac{1}{aq_1^{-1}+b+cq_2}\,q_2^n & \text{pour } n \geqslant 0. \end{cases}$$

Le cas d) est analogue à b).

Si les racines sont multiples : $q_1 = q_2$, en construisant la solution fondamentale bornée on fait appel à la formule

$$u_n = \alpha q_1^n+\beta n q_1^n$$

au lieu de

$$u_n = \alpha q_1^n+\beta q_2^n.$$

Si $|q_1| < 1$, on a pour G_n

$$G_n = \begin{cases} 0 & \text{pour } n \leqslant 0, \\ \dfrac{1}{c}\,nq_1^{n-1} & \text{pour } n \geqslant 0. \end{cases}$$

et si $|q_1| > 1$,

$$G_n = \begin{cases} -\frac{1}{a} n q_1^{n+1} & \text{pour } n \leqslant 0, \\ 0 & \text{pour } n \geqslant 0. \end{cases}$$

Nous avons examiné toutes les possibilités qui peuvent se présenter si $|q_1| \neq 1$, $|q_2| \neq 1$, $a \neq 0$, $c \neq 0$ et constaté l'existence d'une solution fondamentale bornée. Il résulte des formules écrites la décroissance exponentielle de cette solution pour n tendant vers l'infini :

$$|G_n| < G\varrho^{|n|}, \tag{12}$$

$G > 0$ et $0 < \varrho < 1$ étant des constantes.

Ceci étant, ϱ peut être tout nombre vérifiant l'inégalité

$$\varrho > \max\left[\min\left(|q_1|, \frac{1}{|q_1|}\right), \min\left(|q_2|, \frac{1}{|q_2|}\right)\right].$$

Nous avons donc établi l'existence et la forme de la solution fondamentale, i.e. de la solution de l'équation non homogène (9). Si le second membre $\{f_n\}$ de (9) est quelconque, la solution particulière u_n^* se met sous forme de somme de la série

$$u_n^* = \sum_{k=-\infty}^{\infty} G_{n-k} f_k \tag{13}$$

à condition que cette dernière soit convergente. La vérification est calquée sur celle pour l'équation aux différences du premier ordre (voir §2). L'estimation (12) entraîne que la série (13) converge manifestement pour $\{f_k\}$ borné, $|f_k| < F$. Alors

$$|u_n^*| = \left|\sum_{k=-\infty}^{\infty} G_{n-k} f_k\right| \leqslant \sum_{k=-\infty}^{n} |G_{n-k} f_k| + \sum_{k=n+1}^{\infty} |G_{n-k} f_k| \leqslant$$

$$\leqslant GF\left[\sum_{k=-\infty}^{n} \varrho^{n-k} + \sum_{k=n+1}^{\infty} \varrho^{k-n}\right] \leqslant \frac{2G}{1-\varrho} F. \tag{14}$$

Dans le cas de l'équation (9) telle que $|q_1| \neq 1$ et $|q_2| \neq 1$, la solution $\{u_n^*\}$ définie par la formule (13) est la seule solution bornée pour un second membre donné. Dans le cas contraire, nous pourrions obtenir une deuxième solution bornée en ajoutant à celle construite une solution bornée $\{\bar{u}_n\}$ de l'équation homogène (3). Or, les formules donnant la solution générale de cette équation montrent que, lorsque $|q_1| \neq 1$, $|q_2| \neq 1$, il existe une seule solution bornée pour $-\infty < n < \infty$, à savoir $\bar{u}_n \equiv 0$.

En particulier, pour $|q_1| \neq 1$, $|q_2| \neq 1$, il n'y a pas d'autre solution fondamentale bornée que G_n.

Notons que sous les conditions (10) et compte tenu de l'estimation (11), on établit sans peine à partir de (13)

$$|u_n^*| \leqslant \frac{16}{B\theta^2} \sup_m |f_m|. \tag{15}$$

3. Evaluation de la solution fondamentale par les coefficients de l'équation aux différences. Au n° 2 nous avons vu que le comportement de la solution fondamentale G_n de l'équation (9) dépend essentiellement de la position dans le plan complexe des racines q_1 et q_2 de l'équation caractéristique

$$P(q) \equiv a+bq+cq^2 = 0. \tag{4}$$

Un cas particulièrement important pour les applications est celui des a, b, c réels et des racines q_1 et q_2 dont l'une est supérieure et l'autre inférieure en module à l'unité :

$$|q_1| < \varrho, \ |q_2^{-1}| < \varrho, \ 0 < \varrho < 1. \tag{16}$$

Nous allons indiquer un bon critère nécessaire et suffisant d'une telle position des racines qui nous évite leur calcul.

THÉORÈME. *Les racines q_1 et q_2 de l'équation* (4) *à coefficients réels sont l'une supérieure et l'autre inférieure en module à l'unité si et seulement si l'on a une estimation de la forme*

$$\frac{|b|-|a+c|}{|b|+|a|+|c|} \geqslant \theta > 0, \tag{17}$$

θ étant un nombre ; de plus, si (17) *est satisfaite, alors*

$$|q_1| < 1-\frac{\theta}{2}, \quad |q_2^{-1}| < 1-\frac{\theta}{2}. \tag{18}$$

DÉMONSTRATION. Notons que

$$P(1)\cdot P(-1) = (a+c+b)(a+c-b) = |a+c|^2-b^2 = \\ = (|a+c|-|b|)(|a+c|+|b|).$$

Si (17) n'a lieu pour aucun $\theta > 0$, le numérateur de la fraction (17) est soit nul, soit négatif.

Dans le premier cas, $P(1)\cdot P(-1) = 0$, i.e. 1 ou -1 est racine de l'équation (4), et la condition (16) n'est pas remplie.

Dans le second cas, $P(1)\,P(-1) > 0$, i.e. aux points $q = -1$ et $q = 1$, le polynôme $P(q)$ prend des valeurs de même signe. Cela signifie que ou bien il possède sur le segment $-1 \leqslant q \leqslant 1$ deux racines, ou bien il n'en a pas du tout.

Si les racines sont au nombre de deux, l'une et l'autre sont inférieures en module à l'unité (la condition (16) n'est pas remplie), et s'il n'y a pas de racines sur $[-1, 1]$, alors ou bien il n'existe pas de racines réelles (elles sont complexes conjuguées et de même module), ou bien les deux racines sont réelles et de module supérieur à l'unité : la condition (16) n'est pas remplie cette fois non plus.

Si un certain $\theta > 0$ vérifie l'estimation (17), alors $P(-1)\,P(1) < 0$ et les valeurs de $P(q)$ aux extrémités du segment $[-1, 1]$ sont de signes opposés, de sorte que ce segment ne contient qu'une seule racine. L'autre racine,

réelle elle aussi, se situe alors en dehors de $[-1, 1]$ et on a (16) pour un certain $\varrho < 1$. Précisons le dernier résultat, à savoir obtenons l'estimation (18).

Il résulte de (17)

$$|b|-|a+c| \geqslant \theta(|b|+|a|+|c|) > \frac{\theta}{2}|b|+0\cdot|a|+\left[\theta-\left(\frac{\theta}{2}\right)^2\right]\cdot|c|.$$

C'est pourquoi

$$|b|\cdot\left(1-\frac{\theta}{2}\right) > |a+c|+\left[\theta-\left(\frac{\theta}{2}\right)^2\right]\cdot|c| \geqslant$$

$$\geqslant \left|a+c\left\{1-\theta+\left(\frac{\theta}{2}\right)^2\right\}\right| = \left|a+c\left(1-\frac{\theta}{2}\right)^2\right|.$$

Il en découle que les expressions

$$P\left(1-\frac{\theta}{2}\right) = a+c\left(1-\frac{\theta}{2}\right)^2+b\left(1-\frac{\theta}{2}\right),$$

$$P\left[-\left(1-\frac{\theta}{2}\right)\right] = a+c\left(1-\frac{\theta}{2}\right)^2-b\left(1-\frac{\theta}{2}\right)$$

sont de signes différents, de sorte que le polynôme $P(q)$ possède sur le segment $-(1-\theta/2) \leqslant q \leqslant 1-\theta/2$ une racine q_1, $|q_1| < 1-\theta/2$. Il est évident que les nombres

$$q_1' = \frac{1}{q_1}, \quad q_2' = \frac{1}{q_2},$$

inverses des racines de l'équation (4), vérifient l'équation

$$a'+b'q'+c'(q')^2 = 0$$

avec des coefficients $a' = c$, $b' = b$, $c' = a$ tels qu'on ait la même condition (17) :

$$\frac{|b'|-|a'+c'|}{|b'|+|a'|+|c'|} = \frac{|b|-|a+c|}{|b|+|a|+|c|} \geqslant \theta > 0.$$

Aussi l'une des racines q_1', q_2' satisfait à $|q'| < 1-\theta/2$. Cela ne peut être que $q_2' = 1/q_2$, d'où les estimations (18).

Pour une équation à coefficients réels avec la condition (17) on a automatiquement (10) et donc l'estimation (15) de la solution particulière bornée u_n^* de l'équation aux différences non homogène (9).

Exercices

1. Ecrire les solutions générales des équations

$$u_{n-1}-5u_n+6u_{n+1} = 0, \qquad n = 0, \pm 1, \ldots,$$

$$u_{n-1}-\frac{5}{2}u_n+u_{n+1} = 0, \qquad n = 0, \pm 1, \ldots,$$

$$9u_{n-1}+3u_n+u_{n+1} = 0, \qquad n = 0, \pm 1, \ldots$$

2. Trouver la solution bornée pour $n \to +\infty$ de l'équation

$$u_{n-1} - \frac{5}{2} u_n + u_{n+1} = 0$$

qui prend la valeur $u_0 = 1$.

3. Ecrire le millième terme d'une suite $u_0, u_1, u_2, \ldots$ telle que ses deux premiers termes soient égaux à l'unité ($u_0 = 1$, $u_1 = 1$) et les suivants se définissent par la relation de récurrence

$$u_{n+1} = u_{n-1} + u_n, \quad n = 1, 2, \ldots$$

4. Trouver la condition imposée aux racines de l'équation caractéristique qui soit nécessaire et suffisante pour que l'équation aux différences

$$au_{n-1} + bu_n + cu_{n+1} = 0, \quad n = 0, \pm 1, \pm 2, \ldots,$$

admette au moins une solution bornée non triviale (la solution $u_n \equiv 0$ est dite *triviale*).

5. Trouver les conditions imposées aux racines de l'équation caractéristique qui soient nécessaires et suffisantes pour que toutes les solutions de l'équation

$$au_{n-1} + bu_n + cu_{n+1} = 0, \quad n = 0, \pm 1, \ldots,$$

soient bornées.

6. Quelles doivent être les racines de l'équation caractéristique pour que toutes les solutions de l'équation $au_{n-1} + bu_n + cu_{n+1} = 0$ tendent vers 0 pour $n \to +\infty$?

7. Trouver une solution particulière de l'équation aux différences non homogène

$$u_{n-1} - \frac{5}{2} u_n + u_{n+1} = f_n, \quad n = 0, \pm 1, \ldots,$$

dont le second membre est de la forme suivante:

a) $f_n = 1$. INDICATION. Chercher une solution de la forme $u_n^* = A$.
b) $f_n = n$. INDICATION. Chercher une solution de la forme $u_n^* = A + Bn$.
c) $f_n = 3^n$. INDICATION. Chercher une solution de la forme $u_n^* = A \cdot 3^n$.
d) $f_n = \cos n$. INDICATION. Chercher une solution de la forme $u_n^* = A \sin n + B \cos n$.

8. Construire une solution fondamentale bornée de l'équation

$$u_{n-1} + u_n + u_{n+1} = f_n.$$

Possède-t-elle des solutions fondamentales non bornées ?

9. Construire une solution fondamentale de l'équation

$$u_{n-1} - 2u_n + u_{n+1} = f_n.$$

Est-ce qu'il existe une solution fondamentale bornée ?

10. Dans quelle condition imposée aux racines de l'équation caractéristique l'équation aux différences

$$au_{n-1} + bu_n + cu_{n+1} = f_n$$

n'admet-elle pas de solutions fondamentales bornées ?

11. En utilisant la solution fondamentale bornée écrire la solution $(u_0, u_1, \ldots, u_N)$ de l'équation

$$u_{n-1} - \frac{5}{2} u_n + u_{n+1} = f_n, \quad n = 1, 2, \ldots, N-1,$$

qui vérifie les conditions $u_0 = \varphi$, $u_N = \psi$ avec φ et ψ des nombres donnés.

12. Trouver toutes les valeurs propres ϱ et les fonctions propres correspondantes $\psi = \{\psi_m\}$, $m = 0, 1, \ldots, M$, de l'opérateur Λ_{xx}

$$\Lambda_{xx}\psi = \varrho\psi,$$

Λ_{xx} étant un opérateur qui fait correspondre à toute fonction $u = \{u_m\}$ définie sur un réseau une fonction $v = \{v_m\}$ définie sur un réseau d'après les formules

$$v_m = \frac{1}{h^2}(u_{m+1} - 2u_m + u_{m-1}), \quad 0 < m < M,$$

$$v_0 = v_M = 0, \quad Mh = 1.$$

Réponse :

$$\varrho_k = -\frac{4}{h^2}\sin^2\frac{\pi k}{2M}, \quad \psi_m^{(k)} = \sin\frac{k\pi m}{M}, \quad k = 1, 2, \ldots, M-1.$$

CHAPITRE 2

PROBLÈME AUX LIMITES POUR UNE ÉQUATION DU SECOND ORDRE

Les problèmes aux limites du type considéré dans ce chapitre se posent lors de la résolution numérique des équations differentielles ordinaires et des équations aux dérivées partielles à l'aide des schémas aux différences.

§4. Position du problème. Critères de bon conditionnement

1. Position du problème. Le problème aux limites le plus simple consiste à chercher une fonction $\{u_n\}$, $n = 0, 1, \ldots, N$, vérifiant l'équation aux différences

$$a_n u_{n-1} + b_n u_n + c_n u_{n+1} = f_n, \quad n = 1, 2, \ldots, N-1, \tag{1}$$

aux points intérieurs $0 < n < N$ du réseau $0 \leqslant n \leqslant N$ et prenant des valeurs données

$$u_0 = \varphi, \quad u_N = \psi \tag{2}$$

aux extrémités de ce réseau. Le problème aux limites pour un système d'équations aux différences sera énoncé au n°7.

En étudiant l'équation $a_n u_{n-1} + b_n u_n + c_n u_{n+1} = f_n$, $a_n \neq 0$, $c_n \neq 0$, nous avons constaté qu'en se donnant arbitrairement les valeurs de $\{u_n\}$ en deux points successifs quelconques, par exemple pour u_0 et u_1 définis de façon arbitraire, on définit une solution $\{u_n\}$ et une seule.

Il y a intérêt à voir si la solution peut être définie de façon unique en se donnant ses valeurs en deux points pas nécessairement voisins (comme c'est le cas, par exemple, du problème aux limites (1), (2)). On va voir que ce problème peut s'avérer impossible.

Soit le problème aux limites

$$u_{n-1} - u_n + u_{n+1} = 0, \quad n = 1, 2, \ldots, 299, \tag{3}$$

$$u_0 = 0, \quad u_{300} = 1. \tag{4}$$

La solution générale de (3) peut s'écrire (voir §3) :

$$u_n = \gamma_1 \cos \frac{n\pi}{3} + \gamma_2 \sin \frac{n\pi}{3}.$$

La condition $u_0 = 0$ entraîne $\gamma_1 = 0$. Pour avoir $u_{300} = 1$, on doit choisir γ_2 à partir de l'équation

$$u_{300} = \gamma_2 \sin \frac{300\pi}{3} = 1.$$

Or, cette équation n'admet pas de solution puisque, quel que soit γ_2, son premier membre vaut 0 et non pas 1.

Si la condition imposée était $u_{300} = 0$ au lieu de $u_{300} = 1$ (la condition $u_0 = 0$ étant conservée), on aurait toujours $\gamma_1 = 0$ et un γ_2 quelconque :

$$\gamma_2 \sin \frac{300\pi}{3} = \gamma_2 \cdot 0 = 0.$$

Ainsi, le problème aux limites (1), (2) peut en général être impossible ou admettre plus d'une solution. Les problèmes aux limites sont cependant très courants.

Il existe des classes assez vastes d'équations aux différences pour lesquelles le problème (1), (2) admet toujours une solution unique et, de plus, s'avère « peu » sensible aux erreurs d'arrondi résultant de la donnée des seconds membres φ, ψ et $\{f_n\}$, i.e. est « bien conditionné ».

2. Notion de bon conditionnement. En étudiant les schémas aux différences servant à approcher la solution des problèmes aux limites différentiels, on considère en général toute une famille de tels problèmes intervenant avec la diminution du pas du réseau. On considère alors N comme un paramètre dont dépend la famille en question. N croît avec le resserrement du réseau.

Nous dirons que le problème aux limites aux différences (1), (2) à coefficients a_n, b_n, c_n bornés dans leur ensemble, $|a_n|$, $|b_n|$, $|c_n| < M$, est *bien conditionné* s'il admet pour tout N suffisamment grand une et une seule solution $\{u_n\}$ avec des seconds membres φ, ψ et $\{f_n\}$ arbitraires et si les nombres $u_0, u_1, \ldots, u_N$ constituant la solution vérifient l'estimation

$$|u_n| \leqslant M \max \left\{ |\varphi|, |\psi|, \max_m |f_m| \right\}, \tag{5}$$

M étant un nombre indépendant de N.

On range parfois parmi les problèmes bien conditionnés des problèmes pour lesquels on ne peut choisir un M constant, mais tels que M croisse au plus comme une puissance donnée de N, par exemple $M = CN$ ou $M = CN^2$.

La définition du bon conditionnement que nous venons de donner équivaut à l'une de celles adoptées en théorie des systèmes d'équations linéaires : on mesure le conditionnement du système d'équations $Ax = g$ de matrice A par le nombre $\|A\| \cdot \|A^{-1}\|$ qui est le produit de normes des matrices A et A^{-1}.

La validité de l'inégalité (5) signifie que la sensibilité de la solution $\{u_n\}$ aux erreurs (par exemple, aux erreurs de mesure ou à celles d'arrondi) commises lorsqu'on se donnait les seconds membres φ, ψ et $\{f_n\}$ n'augmente pas avec N. En effet, si, au lieu de ces seconds membres, on se donne respecti-

vement $\varphi+\Delta\varphi$, $\psi+\Delta\psi$, $\{f_n+\Delta f_n\}$, la solution $\{u_n\}$ reçoit l'accroissement $\{\Delta u_n\}$, qui est, vu la linéarité du problème (1), (2), solution du problème

$$\left.\begin{aligned} a_n\Delta u_{n-1}+b_n\Delta u_n+c_n\Delta u_{n+1} &= \Delta f_n, \quad 0<n<N, \\ \Delta u_0 = \Delta\varphi, \quad \Delta u_N &= \Delta\psi \end{aligned}\right\}$$

et, en vertu de (5), vérifie l'estimation

$$|\Delta u_n| \leqslant M \max\left\{|\Delta\varphi|, |\Delta\psi|, \max_m |\Delta f_m|\right\}.$$

Tout problème aux limites (1), (2) admettant une solution unique n'est pas bien conditionné, loin de là. Ainsi, si les seconds membres du problème

$$\left.\begin{aligned} u_{n+1}-5u_n+6u_{n-1} &= f_n, \quad 0<n<N, \\ u_0 = \varphi, \quad u_N &= \psi \end{aligned}\right\}$$

reçoivent les accroissements

$$\Delta f_n \equiv 0, \quad \Delta\psi = 0, \quad \Delta\varphi = \varepsilon,$$

la solution $\{u_n\}$ acquiert l'accroissement

$$\Delta u_n = 2^n \frac{1-(^2/_3)^{N-n}}{1-(^2/_3)^N}\Delta\varphi, \quad n = 0, 1, \ldots, N.$$

D'où

$$\Delta u_{N-1} \geqslant 2^{N-1}\cdot\frac{1}{3}\varepsilon.$$

La perturbation ε produite par la donnée de φ a induit sur la solution une perturbation qui augmente rapidement avec N. Il est évident que le nombre M dans l'inégalité (5) doit croître au moins comme l'exponentielle $\frac{1}{3}\cdot 2^{N-1}$.

3. Critère suffisant de bon conditionnement.

THÉORÈME. *Si les coefficients a_n, b_n, c_n remplissent la condition*

$$|b_n| \geqslant |a_n|+|c_n|+\delta, \quad \delta > 0, \tag{6}$$

le problème (1), (2) *est bien conditionné, et la solution $\{u_n\}$ est évaluée par*

$$|u_n| \leqslant \max\left\{|\varphi|, |\psi|, \frac{1}{\delta}\max_m |f_m|\right\}. \tag{7}$$

DÉMONSTRATION. Supposons d'abord que pour φ, ψ et $\{f_n\}$ fixes donnés, le problème (1), (2) possède une solution $\{u_n\}$ et établissons l'estimation (7). Soit $|u_k|$ le plus grand des nombres $|u_n|$, $n = 0, 1, \ldots, N$. Si $k = 0$ ou $k = N$, l'inégalité (7) est évidente puisque $u_0 = \varphi$, $u_N = \psi$. Il nous reste à considérer le cas $0 < k < N$, $|u_k| \geqslant |u_n|$. Compte tenu de (6) nous pou-

vons écrire

$$|b_k|\cdot|u_k| = |-a_k u_{k-1} - c_k u_{k+1} + f_k| \leqslant |a_k|\cdot|u_{k-1}| + \\ + |c_k|\cdot|u_{k+1}| + |f_k| \leqslant (|a_k| + |c_k|)\,|u_k| + |f_k|,$$

$$|u_n| \leqslant |u_k| \leqslant \frac{|f_k|}{|b_k| - |a_k| - |c_k|} \leqslant \frac{|f_k|}{\delta},$$

d'où l'inégalité (6).

Pour achever la démonstration nous devons démontrer que le problème (1), (2) admet une solution et une seule $\{u_n\}$, les seconds membres φ, ψ et $\{f_n\}$ étant arbitraires.

Le problème (1), (2) peut être assimilé à un système de $N+1$ équations linéaires à $N+1$ inconnues $u_0, u_1, \ldots, u_N$. Il faut donc établir que le déterminant de ce système est non nul. On sait du cours d'algèbre que le déterminant d'un système est non nul si et seulement si le système homogène correspondant n'admet pour solution que 0. Pour le système (1), (2), le système homogène s'obtient lorsque $\varphi = \psi = f_m \equiv 0$. L'estimation (7) démontrée pour chaque solution $\{u_n\}$ entraîne qu'on n'a dans ce cas que la solution triviale $u_n \equiv 0$.

Une condition suffisante de bon conditionnement du problème (1), (2) est également

$$\frac{|b_n| - |a_n| - |c_n|}{|b_n| + |a_n| + |c_n|} \geqslant \theta > 0, \quad \max\{|a_n|, |b_n|, |c_n|\} \geqslant B > 0, \tag{8}$$

avec θ et B des constantes indépendantes de N et n.

En effet, (8) entraîne (6) avec la constante

$$\delta = \theta\,(|b_n| + |a_n| + |c_n|) \geqslant \theta B > 0.$$

C'est pourquoi (7) se met sous la forme

$$|u_n| \leqslant \max\left\{|\varphi|, |\psi|, \frac{1}{\theta B}\max_m |f_m|\right\}. \tag{9}$$

4. Critère de bon conditionnement d'un problème aux limites à coefficients constants.

THÉORÈME. *Pour que le problème aux limites*

$$\left.\begin{aligned} au_{n-1} + bu_n + cu_{n+1} &= f_n, \quad 0 < n < N, \\ u_0 = \varphi, \quad u_N &= \psi \end{aligned}\right\} \tag{10}$$

à coefficients constants soit bien conditionné, il faut et il suffit que les racines q_1 et q_2 de l'équation caractéristique

$$a + bq + cq^2 = 0 \tag{11}$$

soient l'une supérieure et l'autre inférieure à l'unité en module, c'est-à-dire

qu'on ait les inégalités de la forme

$$|q_1| \leqslant 1-\frac{\theta}{2}, \quad |q_2^{-1}| \leqslant 1-\frac{\theta}{2}, \tag{12}$$

avec θ une constante positive.

Si les coefficients a, b, c sont réels, le critère de bon conditionnement (12) prend, en vertu de ce que nous avons démontré au n°3 du §3, une forme commode

$$\frac{|b|-|a+c|}{|b|+|a|+|c|} \geqslant \theta > 0. \tag{13}$$

La vérification de (13) est immédiate sans qu'on ait à calculer les racines q_1 et q_2, d'où sa commodité.

Pour la démonstration du critère (12) voir n° 6 du présent paragraphe.

5. Critère de bon conditionnement d'un problème à coefficients variables. Le critère (12) de bon conditionnement d'un problème aux limites lié à une équation aux différences à coefficients constants se généralise au problème

$$a_n u_{n-1}+b_n u_n+c_n u_{n+1} = f_n, \quad 0 < n < N, \tag{1}$$

$$u_0 = \varphi, \quad u_N = \psi \tag{2}$$

à cœfficients variables à condition qu'ils varient d'une façon suffisamment « régulière ». Donnons-en un énoncé rigoureux en supposant que les coefficients de l'équation (1) sont bornés dans leur ensemble : $|a_n| < M$, $|b_n| < M$, $|c_n| < M$, et qu'il n'existe pas de n tel que les trois coefficients a_n, b_n, c_n soient simultanément petits :

$$d_n = \max\{|a_n|, |b_n|, |c_n|\} \geqslant B > 0.$$

M et B sont supposés indépendants de N et n.

THÉORÈME. *Supposons que les coefficients du problème* (1), (2) *vérifient les conditions*

$$\left.\begin{aligned} &|a_k-a_l| \leqslant D\left|\frac{k-l}{N}\right|^{\omega}, \quad |b_k-b_l| \leqslant D\left|\frac{k-l}{N}\right|^{\omega}, \\ &|c_k-c_l| \leqslant D\left|\frac{k-l}{N}\right|^{\omega}, \quad D > 0,\ \omega > 0. \end{aligned}\right\} \tag{14}$$

Pour que le problème considéré soit bien conditionné, il faut et il suffit que les racines q_1 et q_2 de l'équation du deuxième degré

$$a_n+b_n q+c_n q^2 = 0, \quad 0 < n < N, \tag{15}$$

soient telles que

$$|q_1| < 1-\frac{\theta}{2}, \quad |q_2^{-1}| < 1-\frac{\theta}{2}, \tag{16}$$

$\theta > 0$ étant un nombre indépendant de N et n.

Les conditions (14) sont les conditions de régularité des coefficients. Elles sont satisfaites si, par exemple,

$$a_n = a(n/N), \quad b_n = b(n/N), \quad c_n = c(n/N),$$

$a(x)$, $b(x)$, $c(x)$ étant des fonctions définies sur le segment $0 \leqslant x \leqslant 1$ et vérifiant la condition de Hölder :

$$|a(x)-a(x')| \leqslant D\,|x-x'|^{\omega},$$
$$|b(x)-b(x')| \leqslant D\,|x-x'|^{\omega},$$
$$|c(x)-c(x')| \leqslant D\,|x-x'|^{\omega}.$$

L'équation (15) est l'équation caractéristique construite pour l'équation aux différences

$$au_{s-1}+bu_s+cu_{s+1} = 0,$$

où a, b, c sont constants et coïncident avec les valeurs des coefficients variables a_n, b_n, c_n pour n fixe, i.e. $a = a_n$, $b = b_n$, $c = c_n$.

Dans le cas des a_n, b_n, c_n réels, en vertu du n° 3 du §3, on remplace (16) par la condition facile à vérifier

$$\frac{|b_n|-|a_n+c_n|}{|b_n|+|a_n|+|c_n|} \geqslant \theta > 0, \tag{17}$$

où θ ne dépend pas de N et n.

Pour la démonstration du critère (14), (16) ou (14), (17) voir le paragraphe 6. Nous y démontrerons également qu'on ne saurait ignorer les conditions de régularité (14).

Notons que si $|a_n+c_n| = |a_n|+|c_n|$, la condition (17) coïncide avec la condition (8) et assure le bon conditionnement du problème sans assujettir les coefficients à être réguliers et réels.

6. Justification du critère de bon conditionnement d'un problème aux limites à coefficients constants. Démontrons le critère de bon conditionnement du problème aux limites

$$\left.\begin{aligned} au_{n-1}+bu_n+cu_{n+1} &= f_n, \quad 0 < n < N, \\ u_0 = \varphi, \quad u_N &= \psi, \end{aligned}\right\} \tag{10}$$

que nous avons énoncé au n° 4. Nous disons que pour que le problème aux limites (10) soit bien conditionné, il faut et il suffit que les racines de l'équation caractéristique

$$a+bq+cq^2 = 0 \tag{11}$$

satisfassent aux inégalités de la forme

$$|q_1| \leqslant 1-\frac{\theta}{2}, \quad |q_2^{-1}| \leqslant 1-\frac{\theta}{2}, \tag{12}$$

θ étant une constante positive.

Suffisance. Mettons la solution du problème (10) sous forme de somme de deux fonctions définies sur un réseau en posant

$$u_n = \bar{u}_n+\tilde{u}_n, \tag{18}$$

où $\{\bar{u}_n\}$ est solution du problème

$$\left.\begin{aligned} a\bar{u}_{n-1}+b\bar{u}_n+c\bar{u}_{n+1} &= 0, \quad 0 < n < N, \\ \bar{u}_0 = \varphi, \quad \bar{u}_N &= \psi \end{aligned}\right\} \tag{19}$$

et $\{\tilde{u}_n\}$ solution du problème

$$\left.\begin{aligned} a\tilde{u}_{n-1}+b\tilde{u}_n+c\tilde{u}_{n+1} &= f_n, \quad 0 < n < N, \\ \tilde{u}_0 = 0, \quad \tilde{u}_N &= 0. \end{aligned}\right\} \tag{20}$$

La solution du problème (19) est de la forme $\bar{u}_n = Aq_1^n+Bq_2^n$, A et B étant définis par les conditions $\bar{u}_0 = \varphi$, $\bar{u}_N = \psi$:

$$\bar{u}_n = \frac{\varphi-\psi q_2^{-N}}{1-(q_1q_2^{-1})^N}\, q_1^n + \frac{\psi-\varphi q_1^N}{1-(q_1q_2^{-1})^N}\, q_2^{n-N}. \tag{21}$$

En notant $1-\theta/2 = \varrho$, on déduit de (21) :

$$|\bar{u}_n| \leqslant 2\,\frac{\max(\varrho^n, \varrho^{N-n})}{1-\varrho^{2N}} \max(|\varphi|, |\psi|). \tag{22}$$

Quels que soient $N \geqslant 2$ et $n = 0, 1, \ldots, N$, on a donc

$$|\bar{u}_n| \leqslant \frac{2}{1-\varrho} \max(|\varphi|, |\psi|) = \frac{4}{\theta} \max(|\varphi|, |\psi|). \tag{23}$$

Pour N et $N-n$ suffisamment grands, le coefficient de l'inégalité (22) est aussi petit qu'on le veut. Ainsi, lorsque $n > 6/\theta$, $N-n > 6/\theta$, on a

$$\varrho^{6/\theta} = \left[\left(1-\frac{\theta}{2}\right)^{2/\theta}\right]^3 < \left(\frac{4}{9}\right)^2.$$

On a utilisé en l'occurrence l'inégalité connue

$$\left(1-\frac{1}{a}\right)^a\left(1+\frac{1}{b}\right)^b \leqslant \left[\frac{a(1-a^{-1})+b(1+b^{-1})}{a+b}\right]^{a+b} = 1$$

avec $a = 2/\theta$, $b = 2$. Donc

$$\frac{\max(\varrho^n, \varrho^{N-n})}{1-\varrho^{2N}} \leqslant \left(\frac{4}{9}\right)^3 \frac{1}{1-(4/9)^6} < \frac{1}{10}$$

de sorte que (22) fournit pour $n > 6/\theta$, $N-n > 6/\theta$

$$|\bar{u}_n| < \frac{1}{5} \max(|\varphi|, |\psi|). \tag{24}$$

Evaluons la solution $\{\tilde{u}_n\}$ du problème (20). Mettons $\tilde{u}_n$ sous forme de somme

$$\tilde{u}_n = u_n^* + u_n', \quad 0 \leqslant n \leqslant N,$$

des solutions du problème

$$au_{n-1}^*+bu_n^*+cu_{n+1}^* = \begin{cases} f_n, & 0 < n < N, \\ 0, & n \leqslant 0 \text{ ou } n \geqslant N, \end{cases} \tag{25}$$

et du problème

$$\left.\begin{aligned} &au'_{n-1}+bu'_n+cu'_{n+1} = 0, \quad 0 < n < N, \\ &u'_0 = -u^*_0, \quad u'_N = -u^*_N. \end{aligned}\right\} \tag{26}$$

La solution bornée $\{u^*_n\}$ du problème (25) existe, est unique et vérifie l'estimation (15) du §3 :

$$|u^*_n| \leqslant \frac{16}{B\theta^2} \max_m |f_m|, \tag{27}$$

avec $B = \max(|a|, |b|, |c|)$.

En particulier,

$$\left.\begin{aligned} |u^*_0| &\leqslant \frac{16}{B\theta^2} \max_m |f_m|, \\ |u^*_N| &\leqslant \frac{16}{B\theta^2} \max_m |f_m|. \end{aligned}\right\} \tag{27'}$$

On évalue la solution $\{u'_n\}$ du problème (26), qui est un analogue, d'après sa forme, du problème (19), en utilisant la formule (21) et l'estimation (23), où φ et ψ font place respectivement à $-u^*_0$ et $-u^*_N$:

$$|u'_n| \leqslant \frac{4}{\theta} \max(|u^*_0|, |u^*_N|).$$

Tenons également compte de (27') :

$$|u'_n| \leqslant \frac{64}{B\theta^3} \max_m |f_m|. \tag{28}$$

Regroupant les estimations (27) et (28) et vu que $\theta < 2$, on a

$$|\tilde{u}_n| \leqslant \frac{128}{B\theta^3} \max_m |f_m|. \tag{29}$$

Pour la solution $\{u_n\}$ du problème de départ, on obtient donc en réunissant les estimations (23) et (29) :

$$|u_n| \leqslant |\bar{u}_n|+|\tilde{u}_n| \leqslant \frac{128}{B\theta^3} \max_m |f_m|+\frac{4}{\theta} \max(|\varphi|, |\psi|). \tag{30}$$

L'estimation (30) garantit le bon conditionnement $|u_n| \leqslant M \max(|\varphi|, |\psi|, \max_m |f_m|)$, M étant pris égal à

$$M = \frac{128}{B\theta^3}+\frac{4}{\theta}.$$

Si $n > 6/\theta$, $N-n > 6/\theta$, on améliore l'estimation (30) en utilisant non pas (23) mais (24) :

$$|u_n| \leqslant \frac{128}{B\theta^3} \max_m |f_m|+\frac{1}{5} \max(|\varphi|, |\psi|) \tag{31}$$

ou

$$|u_n| < M_1 \max_m |f_m|+\frac{1}{5} \max(|\varphi|, |\psi|), \tag{31'}$$

M_1 dépendant seulement de θ et B et étant indépendant de N. Nous nous servirons de l'estimation (31) au §6.

NÉCESSITÉ. Notons, pour commencer, que si les conditions (12) ne sont remplies pour aucun θ positif, les racines de l'équation caractéristique

$$P(q) \equiv a+bq+cq^2 = 0$$

sont de module inférieur ou supérieur (ou égal pour une racine) à l'unité

$$1)\ |q_1| < \varrho < 1, \quad |q_2| < \varrho < 1, \tag{32}$$

$$2)\ |q_1| > \varrho > 1, \quad |q_2| > \varrho > 1, \tag{33}$$

$$3)\ |q_1| = 1. \tag{34}$$

Montrons que dans les trois cas il n'y a pas de bon conditionnement.

Construisons à cette fin des fonctions $\{u_n\}$ telles qu'elles soient solutions d'un problème de la forme

$$\left.\begin{aligned} au_{n-1}+bu_n+cu_{n+1} &= f_n, \quad 0 < n < N, \\ u_0 = u_N &= 0 \end{aligned}\right\} \tag{35}$$

et qu'on ait les inégalités

$$\max_n |u_n| > M_N \max_m |f_m|, \tag{36}$$

M_N étant une quantité qui croît indéfiniment pour $N \to \infty$.

Considérons le cas (32). Supposons, pour fixer les idées, que $q_1 \neq q_2$ et posons

$$u_n = \begin{cases} q_1^n - q_2^n, & 0 \leqslant n \leqslant N-1, \\ 0, & n = N. \end{cases}$$

Alors

$$\max_n |u_n| \geqslant |u_1| = |q_1 - q_2| > 0. \tag{37}$$

Le second membre $\{f_n\}$ du problème (35) est

$$f_n = au_{n-1}+bu_n+cu_{n+1} = \begin{cases} 0 & \text{si } n \neq N-1, \\ -c(q_1^N - q_2^N) & \text{si } n = N-1. \end{cases}$$

D'où

$$\max_m |f_m| = |f_{N-1}| \leqslant 2|c|\,\varrho^N. \tag{38}$$

Comparons (37) et (38) ; nous voyons qu'on doit poser dans (36)

$$M_N = \frac{|q_1 - q_2|}{2|c|\,\varrho^N} = O\left(\frac{1}{\varrho^N}\right),$$

de sorte que M_N croît exponentiellement avec N.

Le cas (33) est analogue à (32).

Si l'égalité (34) est vraie, on pose

$$u_n = q_1^n \sin\frac{n\pi}{N}, \quad 0 \leqslant n \leqslant N.$$

Evidemment,

$$\max_n |u_n| \geqslant \frac{1}{2}. \tag{39}$$

$|f_n|$ s'évalue par

$$|f_n| = |au_{n-1}+bu_n+cu_{n+1}| = \left|(aq_1^{n-1}+bq_1^n+cq_1^{n+1})\sin\frac{n\pi}{N}+\right.$$

$$\left.+aq_1^{n-1}\left(\sin\frac{(n-1)\pi}{N}-\sin\frac{n\pi}{N}\right)+cq_1^{n+1}\left(\sin\frac{(n+1)\pi}{N}-\sin\frac{n\pi}{N}\right)\right| =$$

$$= \left|aq_1^{n-1}\left(\sin\frac{(n-1)\pi}{N}-\sin\frac{n\pi}{N}\right)+cq_1^{n+1}\left(\sin\frac{(n+1)\pi}{N}-\right.\right.$$

$$\left.\left.-\sin\frac{n\pi}{N}\right)\right| \leqslant (|a|+|c|)\frac{\pi}{N}. \qquad (40)$$

Les inégalités (39) et (40) entraînent (36) si

$$M_N = \frac{N}{2(|a|+|c|)}N.$$

Le problème n'est donc pas bien conditionné si l'on comprend par bon conditionnement l'indépendance de M par rapport à N dans (5).

7. Problèmes aux limites généraux pour des systèmes d'équations aux différences. Le problème (1), (2) est le problème aux limites le plus simple pour une équation du second ordre.

Enonçons (sans démonstration) les conditions nécessaires et suffisantes de bon conditionnement des problèmes aux limites généraux pour des systèmes d'équations aux différences sur un segment maillé (V. Riabenki, Zh. Vychisl. Mat. i Mat. Fiz. 4, 2 (1964)).

Le problème aux limites consiste à obtenir une fonction vectorielle $\{u_n\}$, $n = 0, 1, 2, 3, \ldots, N$, vérifiant les conditions

$$\sum_{k=-k_0}^{k_0} A_{k,n}u_{n+k} = f_n, \quad k_0 \leqslant n \leqslant N-k_0, \qquad (1')$$

$$\sum_{i=0}^{2k_0} \alpha_i u_i = \varphi, \quad \sum_{i=0}^{2k_0} \beta_i u_{N-i} = \psi. \qquad (2')$$

Ici $A_{k,n}$ sont des matrices carrées d'ordre $m \geqslant 1$; u_n, f_n des vecteurs de même dimension; α_i des matrices à m colonnes et $r \geqslant 0$ lignes chacune ; β_i des matries à m colonnes et $s \geqslant 0$ lignes chacune; φ est un vecteur r-dimensionnel donné; ψ un vecteur s-dimensionnel donné.

Le problème (1′), (2′) est bien conditionné s'il admet une solution $\{u_n\}$ pour $\{f_n\}, \varphi, \psi$ arbitraires ; de plus

$$\max_n \|u_n\| \leqslant M \max\left\{\|\varphi\|, \|\psi\|, \max_j \|f_j\|\right\},$$

M étant indépendant de N.

Nous ferons l'hypothèse suivante sur les coefficients $A_{k,n}$:

$$A_{k,n} = A_k\left(\frac{n}{N}\right),$$

avec $A_k(x)$ une matrice définie sur le segment $0 \leqslant x \leqslant 1$ et vérifiant sur ce segment la condition de régularité

$$\|A_k(x)-A_k(x')\| \leqslant D|x-x'|^\omega, \quad D > 0, \ \omega > 0. \qquad (14')$$

Supposons ensuite que

$$d(x) = \max_k \|A_k(x)\| \geqslant B > 0.$$

Avec ces hypothèses restrictives faites, pour que le problème (1′), (2′) soit bien conditionné, il faut et suffit que soient remplies les trois conditions ci-dessous.

1° De toutes les racines μ et ν des équations

$$\det \sum_{k=-k_0}^{k_0} A_k(x)\mu^{k_0+k} = 0,$$

$$\det \sum_{k=-k_0}^{k_0} A_k(x)\nu^{k_0-k} = 0$$

aucune n'est égale à 1 en module, et chaque μ et ν vérifie l'une des quatre inégalités suivantes :

$$|\mu| < 1-\frac{\theta}{2}, \qquad |\nu| < 1-\frac{\theta}{2},$$

$$|\mu^{-1}| < 1-\frac{\theta}{2}, \qquad |\nu^{-1}| < 1-\frac{\theta}{2},$$

où $\theta > 0$ est indépendant de x.

2° La dimension r des matrices α_i est égale au nombre de racines μ de module plus petit que l'unité et la dimension s des matrices β_i au nombre de racines ν de module inférieur à l'unité.

3° Parmi les solutions $\{u_n\}$, $n \geqslant 0$, du problème

$$\left.\begin{aligned} &\sum_{k=-k_0}^{k_0} A_k(0)u_{n+k} = 0, \qquad k_0 \leqslant n < \infty, \\ &\sum_{i=0}^{2k_0} \alpha_i u_i = 0 \end{aligned}\right\}$$

et parmi les solutions $\{u_n\}$, $n \leqslant N$, du problème

$$\left.\begin{aligned} &\sum_{k=-k_0}^{k_0} A_k(1)u_{n+k} = 0, \qquad -\infty < n \leqslant N-k_0, \\ &\sum_{i=0}^{2k_0} \beta_i u_{N-i} = 0 \end{aligned}\right\}$$

il n'y a pas de solutions bornées $\not\equiv 0$.

La dernière condition peut s'écrire comme la non-nullité de certains déterminants dont les éléments sont indépendants de N.

Illustrons le critère énoncé sur l'exemple du problème

$$\left.\begin{aligned} &0\cdot u_{n-1}-2u_n+u_{n+1} = f_n, \qquad 0 < n < N, \\ &u_1-\alpha u_0 = \varphi, \qquad u_N-\beta u_{N-1} = \psi, \end{aligned}\right\}$$

α et β étant des nombres ; $m = 1$, $r = 1$, $s = 1$, $k_0 = 1$. Les racines des équations

$$0-2\mu+\mu^2 = 0 \quad \text{et} \quad 0\cdot\nu^2-2\nu+1 = 0$$

sont

$$\mu_1 = 0, \quad \mu_2 = 2, \quad \nu_1 = {}^1/_2 \quad (\nu_2 = \infty).$$

Aucune d'elles n'est égale à 1 en module, d'où la condition 1°.

La condition 2° est remplie elle aussi. En effet, le nombre de conditions scalaires aux deux extrémités est $r = s = 1$, ce qui est le nombre de racines μ et ν ayant un module plus petit que l'unité.

Etablissons pour quelles valeurs de α le problème

$$\left.\begin{aligned} &0\cdot u_{n-1}-2u_n+u_{n+1} = 0, \qquad n \geqslant 1, \\ &\alpha u_0-u_1 = 0 \end{aligned}\right\}$$

ne possède pas de solutions bornées non triviales. La forme générale de la solution du problème

$$0\cdot u_{n-1}-2u_n+u_{n+1} = 0, \quad n > 0,$$

est

$$u_n = c_1\mu_1^n+c_2\mu_2^n, \quad n \geqslant 0, \quad \mu_1^0 = 1.$$

La condition $|u_n| \leqslant$ const donne $c_2 = 0$. C'est pourquoi

$$u_n = c_1\mu_1^n = \begin{cases} c_1 \text{ si } n = 0, \\ 0 \text{ si } n > 0. \end{cases}$$

Compte tenu de $\alpha u_0-u_1 = 0$, nous voyons que pour $\alpha \neq 0$ il n'y a pas de solutions non triviales et qu'il y en a dans le cas contraire.

Etablissons pour quelles valeurs de β le problème

$$\left.\begin{aligned} 0\cdot u_{n-1}-2u_n+u_{n+1} &= 0, \quad n < N, \\ u_N-\beta u_{N-1} &= 0 \end{aligned}\right\}$$

n'admet pas de solutions non triviales bornées pour $n \to -\infty$. La solution générale du problème $0\cdot u_{n-1}-2u_n+u_{n+1} = 0$, $n < N$, est $u_n = c_1\nu_1^{-n} = c_1(1/2)^{-n} = c_1 2^n$. Cette solution est bornée pour $n \to -\infty$. La condition aux limites $u_N-\beta u_{N-1} = 0$ permet de conclure que

$$c_1 2^N-\beta c_1 2^{N-1} = c_1 2^{N-1}(2-\beta) = 0$$

et la solution non triviale $c_1 \neq 0$, n'existe que pour $\beta = 2$.

Ainsi, le problème aux limites considéré est bien conditionné quels que soient $\alpha \neq 0$ et $\beta \neq 2$. Si $\alpha = 0$ et $\beta = 2$, le bon conditionnement n'a pas lieu.

Exercices

Le problème aux limites aux différences

$$\left.\begin{aligned} au_{n-1}+bu_n+cu_{n+1} &= f_n, \quad 0 < n < N, \\ u_0-\alpha u_1 &= \varphi, \quad u_{N-1}-\beta u_N = \psi \end{aligned}\right\} \qquad (*)$$

sera dit *bien conditionné* s'il n'admet pour chaque N qu'une solution et une seule et si les nombres $u_0, u_1, \ldots, u_N$ constituant la solution $\{u_n\}$ vérifient l'inégalité $|u_n| < M\max(|\varphi|, |\psi|, \max_m |f_m|)$, M étant indépendant de N.

1. Si l'équation caractéristique $a+bq+cq^2 = 0$ a ses deux racines q_1, q_2 inférieures (resp. supérieures) à 1 en module, le problème (*) ne peut pas être bien conditionné. Démontrer. Supposer, pour simplifier, $q_1 \neq q_2$.

2. Si au moins l'une des racines q_1, q_2 de l'équation caractéristique a un module égal à l'unité, le problème (*) ne peut pas être bien conditionné. Démontrer.

3. Si $|q_1| < 1$, $|q_2| > 1$, mais

$$1-\alpha q_1 = 0 \quad \text{ou} \quad 1-\beta q_2 = 0,$$

le problème (*) ne peut pas être bien conditionné. Démontrer.

4. Le problème (*) est bien conditionné si et seulement si $|q_1| < 1$, $|q_2| > 1$ et si $1-\alpha q_1 \neq 0$, $1-\beta q_2 \neq 0$. Démontrer.

5. Le problème à coefficients constants (complexes)

$$au_{n-1}+bu_n+cu_{n+1} = f_n, \quad n = 0, \pm 1, \ldots,$$

et avec un second membre périodique quelconque

$$f_{n+N} = f_n$$

admet pour tout N une solution périodique $\{u_n\}$, $u_{n+N} = u_n$, évaluée par

$$|u_n| \leqslant M \max_m |f_m|,$$

M ne dépendant pas de N et de $\{f_n\}$ si et seulement si aucune des racines de l'équation caractéristique $a+bq+cq^2 = 0$ n'est égale à l'unité en module. Démontrer.

§5. Balayage

1. Description. Décrivons une méthode simple et commode de résolution du problème aux limites aux différences de la forme examinée au §4 :

$$\left.\begin{aligned} a_n u_{n-1}+b_n u_n+c_n u_{n+1} &= f_n, \quad 0 < n < N, \\ u_0 = \varphi, \quad u_N &= \psi. \end{aligned}\right\} \tag{1}$$

Il s'agit d'une variante de la méthode d'élimination des inconnues, qui s'appelle *méthode du balayage.*

Ecrivons l'équation $u_0 = \varphi$ du système (1) sous la forme

$$u_0 = L_{1/2}u_1+K_{1/2},$$

où $L_{1/2} = 0$ et $K_{1/2} = \varphi$. Eliminons u_1 de l'équation

$$a_1u_0+b_1u_1+c_1u_2 = f_1$$

correspondant à $n = 1$ dans le système (1) et utilisons à cet effet l'égalité $u_0 = L_{1/2}u_1+K_{1/2}$. Ecrivons le résultat sous une forme résolue en u_1 :

$$u_1 = L_{3/2}u_2+K_{3/2}$$

en notant

$$L_{3/2} = \frac{-c_1}{b_1}, \quad K_{3/2} = \frac{a_1\varphi-f_1}{-b_1}.$$

Utilisons $u_1 = L_{3/2}u_2+K_{3/2}$ pour éliminer u_1 de l'équation

$$a_2u_1+b_2u_2+c_2u_3 = f_2$$

correspondant à $n = 2$. En explicitant on a

$$u_2 = L_{5/2}u_3+K_{5/2}.$$

On peut poursuivre le processus d'élimination décrit pour $n = 3, 4, \ldots$

Portons

$$u_{n-1} = L_{n-1/2}u_n+K_{n-1/2}$$

dans l'équation

$$a_nu_{n-1}+b_nu_n+c_nu_{n+1} = f_n,$$

il vient

$$u_n = \frac{-c_n}{b_n+a_nL_{n-1/2}}u_{n+1}+\frac{f_n-a_nK_{n-1/2}}{b_n+a_nL_{n-1/2}}.$$

Les coefficients des relations

$$u_n = L_{n+1/2}u_{n+1}+K_{n+1/2}$$

obtenues par élimination se calculent donc par les formules de récurrence

$$\left.\begin{aligned} L_{n+1/2} &= \frac{-c_n}{b_n+a_nL_{n-1/2}}, \\ K_{n+1/2} &= \frac{f_n-a_nK_{n-1/2}}{b_n+a_nL_{n-1/2}}. \end{aligned}\right\} \tag{2}$$

La dernière des relations ainsi obtenues est de la forme

$$u_{N-1} = L_{N-1/2}u_N+K_{N-1/2}.$$

Comme $u_N = \psi$, il est possible de calculer u_{N-1} :

$$u_{N-1} = L_{N-1/2}\psi+K_{N-1/2}.$$

u_{N-2}, u_{N-3} et suivants sont définis par les égalités

$$u_{N-2} = L_{N-3/2}u_{N-1}+K_{N-3/2},$$
$$u_{N-3} = L_{N-5/2}u_{N-2}+K_{N-5/2},$$

et ainsi de suite, et le processus s'arrête avec u_1.

Résumons en bref l'essence de la procédure décrite.

On calcule les coefficients $L_{n+1/2}$, $K_{n+1/2}$ dans l'ordre de croissance des numéros (balayage en sens direct) à l'aide des récurrences (2), $L_{1/2} = 0$ et $K_{1/2} = \varphi$ étant donnés.

Puis, toujours par récurrence, on calcule les inconnues u_n dans l'ordre de décroissance des numéros (balayage en sens inverse ou par remontée) d'après les formules

$$\left.\begin{aligned} &u_N = \psi, \\ &u_n = L_{n+1/2}u_{n+1}+K_{n+1/2}, \quad n = N-1, N-2, \ldots, 1. \end{aligned}\right\} \tag{3}$$

Notons que le calcul par balayage de la solution $u_0, u_1, \ldots, u_N$ du système (1) à $N+1$ équations exige des opérations arithmétiques excédant un nombre de fois fini le nombre d'inconnues. La résolution d'un système linéaire de N équations à N inconnues par élimination coûte généralement des opérations arithmétiques en nombre de l'ordre de N^3. Or, en utilisant les caractères spécifiques du système (1) on a réussi à le résoudre par balayage au bout d'un nombre moins important d'opérations arithmétiques.

Nous allons montrer plus bas (§7) qu'en résolvant par la méthode du balayage le problème aux limites (1) vérifiant l'une des conditions de bon conditionnement indiquées au §4 :

$$|b_n| \geqslant |a_n|+|c_n|+\delta, \quad \delta > 0. \tag{4}$$

ou

$$\frac{|b_n|-|a_n|-|c_n|}{|b_n|+|a_n|+|c_n|} \geqslant \theta > 0, \quad d_n = \max(|a_n|, |b_n|, |c_n|) \geqslant B > 0,$$

ou encore

$$\frac{|b_n|-|a_n+c_n|}{|b_n|+|a_n|+|c_n|} \geqslant \theta > 0, \quad d_n \geqslant B > 0,$$

$$|a_k-a_l| \leqslant D\left|\frac{k-l}{N}\right|^{\omega}, \quad |b_k-b_l| \leqslant D\left|\frac{k-l}{N}\right|^{\omega},$$

$$|c_k-c_l| \leqslant D\left|\frac{k-l}{N}\right|^{\omega}, \quad D > 0, \omega > 0$$

les expressions $b_n+a_nL_{n-1/2}$ par lesquelles on divise ne s'annulent pas et les erreurs de calcul ne s'accumulent pas et ne conduisent pas à des erreurs sur la solution qui croissent avec N.

Ces deux propriétés remarquables de la méthode du balayage — un nombre faible d'opérations arithmétiques qu'elle coûte et le fait qu'elle est peu sensible aux erreurs de calcul — en font un algorithme fort commode.

2. Exemple d'algorithme numériquement instable. Le problème aux différences bien conditionné (1) se résout par divers algorithmes. L'algorithme de balayage que nous venons de décrire a l'avantage de coûter peu d'opérations arithmétiques et d'être numériquement stable. Voici un algorithme plus simple mais qui est numériquement instable et donc impuissant pour N importants.

Après avoir défini $U_0^{(1)} = \varphi$, $U_1^{(1)} = 0$, cherchons la solution $U^{(1} = \{U_n^{(1)}\}$, $n = 0, 1, \ldots, N$, de l'équation aux différences (1). Visiblement, on a en général $U_N^{(1)} \neq \psi$. On se donne $U_0^{(2)} = \varphi$, $U_1^{(2)} = 1$ et on calcule la solution $U^{(2)} = \{U_n^{(2)}\}$ qui ne satisfait pas elle non plus à la condition à l'extrémité droite. Posons

$$u_n = \sigma U_n^{(1)}+(1-\sigma)U_n^{(2)}, \quad n = 0, 1, \ldots N. \tag{5}$$

Il est évident que, quel que soit σ, on a la condition $u_0 = \varphi$ et l'équation (1). Choisissons un σ tel que

$$u_N = \sigma U_N^{(1)}+(1-\sigma)U_N^{(2)} = \psi,$$

i.e. posons

$$\sigma = \frac{\psi - U_N^{(2)}}{U_N^{(1)} - U_N^{(2)}}$$

et obtenons moyennant la formule (5) la solution cherchée du problème (1).

Dans le cas d'une machine fictive et des calculs exacts, notre algorithme ferait parfaitement l'affaire. Montrons maintenant que sa sensibilité aux erreurs d'arrondi pour le problème bien conditionné (1) augmente rapidement pour $N \to \infty$. Plaçons-nous dans le cas $a_n \equiv 1$, $b_n \equiv -26/5$, $c_n \equiv 1$, $f_n \equiv 0$.

La condition (4) de bon conditionnement étant remplie, la solution exacte du problème aux limites aux différences est donnée par

$$u_n = \frac{5^{N-n}-5^{n-N}}{5^N-5^{-N}}\varphi + \frac{5^n-5^{-n}}{5^N-5^{-N}}\psi. \tag{6}$$

En vertu de (5) §3, on a pour $U_n^{(1)}$, $U_n^{(2)}$:

$$U^{(1)} = -\frac{\varphi}{24}5^n+\frac{\varphi}{24}5^{2-n},$$

$$U_n^{(2)} = \frac{5-\varphi}{24}5^n+\left[5-\frac{25}{24}(5-\varphi)\right]5^{-n}.$$

Du moment que les valeurs $\max\limits_n |U_n^{(1)}|$ et $\max\limits_n |U_n^{(2)}|$ croissent comme 5^N, les nombres obtenus pour $U_n^{(1)}$, $U_n^{(2)}$, N étant importants, dépassent la capacité de la machine. Supposons qu'il n'en est rien et que $\{U_n^{(1)}\}$, $\{U_n^{(2)}\}$ et σ sont calculés exactement. Supposons de plus que l'unique erreur d'arrondi commise ε l'est dans le calcul de $1-\sigma$. Conformément à la formule (5) nous obtenons alors, au lieu de $\{u_n\}$,

$$\{u_n+\Delta u_n\},$$

avec $\Delta u_n = \varepsilon U_n^{(2)}$.

L'erreur $\{\Delta u_n\}$ pour $n \sim N$ est de la forme

$$\Delta u_n \sim 5^N\varepsilon$$

et, pour l'erreur relative fixe ε commise en calculant $1-\sigma$, elle s'accroît rapidement « en écrasant » la solution exacte $\{u_n\}$ qui reste bornée vu la formule (6).

L'algorithme décrit porte le nom de *méthode du tir*. Des cas peuvent se présenter (voir §20) où il s'avère stable et efficace.

Exercices

1. Comment modifier l'algorithme de balayage pour calculer la solution $\{u_n\}$, $0 \leqslant n \leqslant N$, de l'équation aux différences

$$a_nu_{n-1}+b_nu_n+c_nu_{n+1} = f_n, \quad 0 < n < N,$$

avec les conditions aux limites de la forme

$$u_0-\alpha u_1 = \varphi, \quad u_N-\beta u_{N-1} = \psi$$

si α et β sont non nuls ?

2. En calculant la solution du problème

$$\left.\begin{aligned} a_nu_{n-1}+b_nu_n+c_nu_{n+1} &= f_n, \quad 0 < n < N\}, \\ u_0 = \varphi, \quad u_N &= \psi \end{aligned}\right\}$$

on pourrait éliminer les inconnues u_n dans l'ordre de décroissance des numéros n.

Ecrire les récurrences permettant de calculer les coefficients $L_{n+1/2}$, $K_{n+1/2}$ des relations de balayage

$$u_{n+1} = L_{n+1/2}u_n+K_{n+1/2}, \quad n = N-1, N-2, \ldots, 0.$$

3. Les coefficients a_n, b_n, c_n de l'équation aux différences étant astreints aux contraintes $a_n > 0$, $c_n > 0$, $-b_n > a_n+c_n+\delta$, montrer que les coefficients de balayage $L_{n-1/2}$ apparus

en résolvant le problème

$$\left.\begin{aligned} u_0 &= \alpha u_1+\varphi, \quad 0<\alpha<1, \\ a_n u_{n-1}+b_n u_n+c_n u_{n+1} &= f_n, \quad 0<n<N, \\ u_{N-1} &= \beta u_N+\psi \end{aligned}\right\}$$

vérifient les inégalités $0 \leqslant L_{n-1/2} \leqslant 1$. L'influence de ce fait sur l'accumulation d'erreurs dans le balayage par remontée ? Est-ce que le dénominateur des formules régissant le balayage en sens direct peut devenir nul ?

4. Quelle variante de la méthode du balayage choisir pour calculer la solution du problème précédent si $\alpha = 10$, $\beta = -0{,}5$? Tenir compte de la nécessité d'une division par zéro pour obtenir les coefficients des relations de balayage par les formules de récurrence.

CHAPITRE 3

JUSTIFICATION DE LA MÉTHODE DU BALAYAGE (*)

§6. Propriétés des problèmes aux limites bien conditionnés

Nous allons démontrer le critère de bon conditionnement d'un problème aux limites aux différences de la forme

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} = f_n, \quad 0 < n < N, \\ u_0 = \varphi, \quad u_N = \psi, \end{aligned}\right\} \tag{1}$$

critère énoncé au n° 5 §4, et établir certaines propriétés des problèmes aux limites aux différences bien conditionnés, qui nous serviront au §7 à justifier l'algorithme de balayage.

1. Evaluation des solutions d'un problème aux limites à coefficients perturbés. Considérons un problème du type (1) :

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} = f_n, \quad p < n < q, \\ u_p = \varphi, \quad u_q = \psi, \end{aligned}\right\} \tag{1'}$$

avec p et $q \geqslant p+2$ des entiers. Le fait de numéroter les composantes de la solution de p à q et non de 0 à N est sans importance encore que commode pour l'exposé ultérieur. Les coefficients seront supposés bornés dans leur ensemble : $|a_n|$, $|b_n|$, $|c_n| < M_1$, M_1 ne dépendant pas de N et n.

Admettons que le problème (1') soit résoluble pour φ, ψ et $\{f_n\}$ arbitraires, les nombres u_p, u_{p+1}, ..., u_q, composantes de la solution, vérifiant les inégalités

$$|u_n| \leqslant M_1 \max_m |f_m| + M_2 \max(|\varphi|, |\psi|), \tag{2}$$

où M_1 et M_2 sont des constantes positives, $M_1 \geqslant M_2$, $M_1 \geqslant 1$.

Considérons le problème

$$\left.\begin{aligned} \tilde{a}_n \tilde{u}_{n-1} + \tilde{b}_n \tilde{u}_n + \tilde{c}_n \tilde{u}_{n+1} = f_n, \quad p < n < q, \\ \tilde{u}_p = \varphi, \quad \tilde{u}_q = \psi. \end{aligned}\right\} \tag{3}$$

(*) La matière du chap. 3 n'étant pas utilisée dans l'exposé qui suit, ce chapitre peut être omis en première lecture.

Sous l'hypothèse des coefficients recevant une perturbation $\tilde{a}_n - a_n$, $\tilde{b}_n - b_n$, $\tilde{c}_n - c_n$ pas très importante, à savoir

$$\left.\begin{aligned} |\tilde{a}_n - a_n| &< \varepsilon < \frac{1}{6M_1}, \\ |\tilde{b}_n - b_n| &< \varepsilon < \frac{1}{6M_1}, \\ |\tilde{c}_n - c_n| &< \varepsilon < \frac{1}{6M_1}, \end{aligned}\right\} \tag{4}$$

le système perturbé (3) jouit des quatre propriétés suivantes :

1° Le problème (3) admet une solution $\{\tilde{u}_n\}$ pour n'importe quel second membre.

2° La solution $\{\tilde{u}_n\}$ vérifie une estimation de la forme (2), où M_1 et M_2 sont remplacés respectivement par $2M_1$ et $2M_2$:

$$|\tilde{u}_n| \leqslant 2M_1 \max_m |f_m| + 2M_2 \max(|\varphi|, |\psi|). \tag{5}$$

3° Les cœfficients $\tilde{a}_n$, $\tilde{b}_n$, $\tilde{c}_n$ sont évalués par

$$|\tilde{a}_n| < M_1 + \frac{1}{6M_1}, \quad |\tilde{b}_n| < M_1 + \frac{1}{6M_1}, \quad |\tilde{c}_n| < M_1 + \frac{1}{6M_1}.$$

4° Les solutions $\{u_n\}$ et $\{\tilde{u}_n\}$ diffèrent d'une quantité peu importante, à savoir

$$|\tilde{u}_n - u_n| \leqslant \varepsilon\left[6M_1^2 \max_m |f_m| + 6M_1M_2 \max(|\varphi|, |\psi|)\right]. \tag{6}$$

La propriété 3° est évidente. Démontrons 2° et en déduisons 1°. Supposons le système (3) possible pour certains seconds membres. Fixant ceux-ci, notons

$$\mu = \max_k |\tilde{u}_k|$$

et obtenons pour μ l'inégalité

$$\mu \leqslant 2M_1 \max_m |f_m| + 2M_2 \max(|\varphi|, |\psi|). \tag{7}$$

A cette fin, mettons (3) sous la forme

$$\left.\begin{aligned} a_n\tilde{u}_{n-1} + b_n\tilde{u}_n + c_n\tilde{u}_{n+1} &= f_n + (a_n - \tilde{a}_n)\tilde{u}_{n-1} + (b_n - \tilde{b}_n)\tilde{u}_n + \\ &\quad + (c_n - \tilde{c}_n)\tilde{u}_{n+1}, \quad 0 < n < N, \\ \tilde{u}_0 &= \varphi, \quad \tilde{u}_N = \psi. \end{aligned}\right\} \tag{8}$$

Cette écriture et les estimations (2) et (4) entraînent

$$\mu \leqslant M_1\left(\max_m |f_m| + \frac{3}{6M_1}\mu\right) + M_2 \max(|\varphi|, |\psi|) \leqslant$$
$$\leqslant \frac{1}{2}\mu + M_1 \max_m |f_m| + M_2 \max(|\varphi|, |\psi|).$$

Résolvant la dernière inégalité par rapport à μ nous obtenons (7), d'où (5).

Il découle de l'inégalité (5) que le système homogène correspondant au problème (3) et intervenant pour $\varphi = \psi = f_n \equiv 0$ n'a pas d'autre solution que $\tilde{u}_n \equiv 0$. Aussi le déterminant du système (3) est non nul et le problème (3) admet une solution unique pour des seconds membres quelconques. Nous avons donc démontré les propriétés 1° et 2°. Il nous reste à démontrer 4°, i.e. l'inégalité (6).

Retranchons terme à terme les égalités (1) des égalités (8), il vient

$$\left.\begin{aligned} a_n(\tilde{u}_{n-1}-u_{n-1})+b_n(\tilde{u}_n-u_n)+c_n(\tilde{u}_{n+1}-u_{n+1}) = \\ = (a_n-\tilde{a}_n)\tilde{u}_{n-1}+(b_n-\tilde{b}_n)\tilde{u}_n+(c_n-\tilde{c}_n)\tilde{u}_{n+1}, \quad 0 < n < N, \\ \tilde{u}_0-u_0 = 0, \ \tilde{u}_N-u_N = 0. \end{aligned}\right\}$$

Appliquons (2) :

$$|\tilde{u}_n-u_n| \leqslant M_1 \max_m |(\tilde{a}_m-a_m)\tilde{u}_{m-1}+(\tilde{b}_m-b_m)\tilde{u}_m+(\tilde{c}_m-c_m)\tilde{u}_{m+1}|.$$

Compte tenu de (4) et (5) nous en déduisons

$$|\tilde{u}_n-u_n| \leqslant M_1\varepsilon\left[3\cdot 2M_1 \max_m |f_m|+3\cdot 2M_2 \max(|\varphi|, |\psi|)\right],$$

i.e. l'inégalité (6).

Considérons maintenant le problème produit par une perturbation des coefficients et des seconds membres de (1') :

$$\left.\begin{aligned} \tilde{a}_n\tilde{u}_{n-1}+\tilde{b}_n\tilde{u}_n+\tilde{c}_n\tilde{u}_{n+1} = \tilde{f}_n, \quad p < n < q, \\ \tilde{u}_p = \tilde{\varphi}, \quad \tilde{u}_q = \tilde{\psi}. \end{aligned}\right\} \tag{9}$$

On montre que

$$\begin{aligned} |\tilde{u}_n-u_n| \leqslant \varepsilon\left[[6M_1^2 \max_m |\tilde{f}_m|+6M_1M_2 \max(|\tilde{\varphi}|, |\tilde{\psi}|)\right]+ \\ +M_2 \max(|\tilde{\varphi}-\varphi|, |\tilde{\psi}-\psi|)+M_1 \max_m |\tilde{f}_m-f_m|. \end{aligned} \tag{10}$$

La démonstration d'après le schéma suivant se fait aisément.

Modifions les seconds membres en conservant les coefficients. Moyennant (2) nous constatons que chaque u_n varie d'au plus

$$M_1 \max_m |\tilde{f}_m-f_m|+M_2 \max(|\tilde{\varphi}-\varphi|, |\tilde{\psi}-\psi|).$$

Perturbons ensuite les coefficients du système ainsi modifié. Nous constatons qu'en vertu de la propriété 4°, les composantes u_n reçoivent une perturbation supplémentaire ne dépassant pas

$$\varepsilon\left[6M_1^2 \max_m |\tilde{f}_m|+6M_1M_2 \max(|\tilde{\varphi}|, |\tilde{\psi}|)\right],$$

ce qui conduit à l'estimation (10).

Forts des conséquences décrites de l'inégalité (2), nous en déduisons encore une : supposons que pour un certain $\lambda > 0$, $p+\lambda < n < q-\lambda$, les solutions du système (1') sont évaluées par

$$|u_n| < M_1 \max_m |f_m|+M_2' \max_m (|\varphi|, |\psi|).$$

La solution du système perturbé

$$\left.\begin{aligned} \tilde{a}_n\tilde{u}_{n-1}+\tilde{b}_n\tilde{u}_n+\tilde{c}_n\tilde{u}_{n+1} &= f_n, \quad p < n < q, \\ \tilde{u}_p = \varphi, \quad \tilde{u}_q &= \psi, \end{aligned}\right\}$$

vérifiant les conditions

$$|\tilde{a}_n-a_n|,\ |\tilde{b}_n-b_n|,\ |\tilde{c}_n-c_n| < \varepsilon < \frac{1}{24M_1^2} < \frac{1}{6M_1}, \tag{11}$$

satisfait, dans ces mêmes conditions $p+\lambda < n < q-\lambda$, à l'inégalité

$$|\tilde{u}_n| \leqslant 2M_1 \max_m |f_m| + \left(M_2'+\frac{1}{4}\right) \max(|\varphi|, |\psi|). \tag{12}$$

Pour nous en assurer, définissons une fonction discrète auxiliaire $\{v_n\}$ comme solution du système

$$\left.\begin{aligned} a_nv_{n-1}+b_nv_n+c_nv_{n+1} &= 0, \quad p < n < q, \\ v_p = \varphi, \quad v_q &= \psi. \end{aligned}\right\}$$

Lorsque $p+\lambda < n < q-\lambda$, on a

$$|v_n| < M_2' \max(|\varphi|, |\psi|). \tag{13}$$

Evaluons $|\tilde{u}_n-v_n|$ moyennant l'inégalité (10), d'où il résulte, compte tenu de (11),

$$\begin{aligned} |\tilde{u}_n - v_n| &\leqslant \varepsilon\left[6M_1^2 \max_m |f_m| + 6M_1M_2 \max(|\varphi|, |\psi|)\right] + \\ &+ M_1 \max_m |f_m| \leqslant \frac{1}{4} \max(|\varphi|, |\psi|) + (1+M_1) \max_m |f_m| \leqslant \\ &\leqslant \frac{1}{4} \max(|\varphi|, |\psi|) + 2M_1 \max_m |f_m|. \end{aligned}$$

Vu l'estimation (13), il en découle de suite l'inégalité (12).

REMARQUE. Soulignons que la quantité ε de (4) limitant la perturbation des coefficients du problème initial de façon que celui-ci reste possible, ainsi que les coefficients de l'estimation (5) de la solution du problème perturbé et ceux des inégalités (6) et (10) évaluant l'écart entre la solution du problème perturbé et celle du problème de départ, dépend uniquement des coefficients M_1 et M_2 de l'estimation (2). Les valeurs effectives des coefficients de l'équation aux différences et le nombre $q-p+1$ de points ne sont pas importants en eux-mêmes : ils n'influent que par le truchement des constantes M_1 et M_2 vérifiant l'inégalité (2).

2. Démonstration du critère de bon conditionnement. Au n° 5 §4 nous avons formulé le critère de bon conditionnement du problème (1), les coefficients étant supposés réguliers :

$$\left.\begin{aligned} |a_k-a_l| &\leqslant D\left|\frac{k-l}{N}\right|^{\omega}, \quad |b_k-b_l| \leqslant D\left|\frac{k-l}{N}\right|^{\omega}, \\ |c_k-c_l| &\leqslant D\left|\frac{k-l}{N}\right|^{\omega}, \quad D > 0, \quad \omega > 0, \end{aligned}\right\} \tag{14}$$

dans les conditions

$$\left.\begin{aligned} d_n &= \max(|a_n|, |b_n|, |c_n|) \geqslant B > 0, \\ |a_n| &\leqslant M_1, \quad |b_n| \leqslant M_1, \quad |c_n| \leqslant M_1. \end{aligned}\right\} \tag{14'}$$

Sous les hypothèses (14) et (14'), le problème (1) est bien conditionné si et seulement si les racines de l'équation du deuxième degré

$$a_n + b_n q + c_n q^2 = 0$$

vérifient les inégalités

$$|q_1| \leqslant 1 - \frac{\theta}{2}, \quad |q_2^{-1}| \leqslant 1 - \frac{\theta}{2}, \tag{15}$$

avec $\theta > 0$ indépendant de N et n.

La *nécessité* se démontre à peu près comme on l'a fait au n° 6 §4 pour le cas des coefficients constants. Nous passerons donc outre.

On démontre la *suffisance* en recourant au critère de bon conditionnement (15) (voir n° 4 §4) du problème aux limites aux différences

$$\left.\begin{aligned} au_{n-1} + bu_n + cu_{n+1} &= f_n, \quad p < n < q, \\ u_p = \varphi, \quad u_q &= \psi \end{aligned}\right\} \tag{16}$$

à coefficients constants, où p et q, $q \geqslant p+2$, sont des entiers quelconques. A la différence du §4, nous numéroterons les composantes de la solution $\{u_n\}$ non pas par $n = 0, 1, \ldots, N$, mais par $n = p, p+1, \ldots, q$, ce qui est d'ailleurs sans importance. Le problème (16) admet toujours une solution, de plus, quels que soient n, $p \leqslant n \leqslant q$, on a l'estimation (30) §4 :

$$|u_n| \leqslant M_1 \max_m |f_m| + M_2 \max(|\varphi|, |\psi|), \quad p \leqslant n \leqslant q, \tag{17}$$

et pour n, $p + \frac{6}{\theta} < n < q - \frac{6}{\theta}$, l'estimation (31) §4 :

$$|u_n| \leqslant M_1 \max_m |f_m| + M_2' \max(|\varphi|, |\psi|), \tag{18}$$

où

$$M_1 = \frac{128}{B\theta^3}, \quad M_2 = \frac{4}{\theta}, \quad M_2' = \frac{1}{5}.$$

Choisissons ε en posant

$$\varepsilon = \frac{1}{24 M_1^2}. \tag{19}$$

N sera supposé suffisamment grand pour qu'on ait

$$D\left|\frac{24}{N\theta}\right|^{\omega} < \varepsilon, \quad \text{i.e.} \quad N > \frac{24}{\theta}\left|\frac{D}{\varepsilon}\right|^{1/\omega}. \tag{20}$$

Passons à la démonstration du bon conditionnement du problème (1). Considérons à cette fin un problème aux limites de la forme

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, p < n < q, \\ u_p = \varphi, \quad u_q &= \psi, \end{aligned}\right\} \tag{21}$$

p et q étant des nombres fixes quelconques, $0 \leqslant p$, $q \leqslant N$, $q \geqslant p+2$. Dans le cas particulier $p = 0$, $q = N$, ce problème coïncide avec (1), et il s'obtient en général de ce dernier par « troncature », à savoir en éliminant les équations pour $n \leqslant p$ et $n \geqslant q$ et en définissant u_p et u_q. Montrons que pour un N quelconque vérifiant la condition (20) le problème (21) n'admet qu'une seule solution pour des seconds membres arbitraires et que les nombres $\{u_n\}$, $p \leqslant n \leqslant q$, satisfont à une estimation de la forme

$$|u_n| \leqslant M \max\left(|\varphi|, |\psi|, \max_m |f_m|\right), \tag{22}$$

avec M une constante dépendant de B, θ et indépendante de N, p, q.

Discutons les cas $q-p \leqslant 24/\theta$ et $q-p > 24/\theta$.

Si $q-p \leqslant 24/\theta$, les coefficients du problème (21), où k et l sont quelconques, $p < k$, $l < q$, vérifient, en vertu des conditions de régularité (14) et vu que N est, conformément à (20), suffisamment grand, les estimations suivantes :

$$|a_k - a_l| \leqslant D\left|\frac{k-l}{N}\right|^{\omega} \leqslant D\left|\frac{q-p}{N}\right|^{\omega} \leqslant D\left|\frac{24}{\theta N}\right|^{\omega} < \varepsilon,$$
$$|b_k - b_l| < \varepsilon, \quad |c_k - c_l| < \varepsilon.$$

Ces coefficients sont « presque » constants et ils s'écartent d'au plus ε des coefficients du problème (16), où on a pris pour a, b, c respectivement a_{p+1}, b_{p+1}, c_{p+1}. La solution du problème (16) est évaluée par (17). Le nombre ε est choisi d'après la formule (19) conformément à la condition (4). La solution de (21) peut donc être évaluée par l'inégalité (5) :

$$|u_n| \leqslant \frac{256}{B\theta^5} \max_m |f_m| + \frac{8}{\theta} \max(|\varphi|, |\psi|). \tag{23}$$

Plaçons-nous maintenant dans le cas $q-p > 24/\theta$, en particulier $p = 0$, $q = N$. Supposons que pour certains φ, ψ et $\{f_m\}$ fixes il existe une solution $\{u_n\}$, $p \leqslant n \leqslant q$. Formons une suite d'entiers $p = N_0 < N_1 < \ldots < N_r = q$ telle qu'on ait

$$\frac{6}{\theta} < N_{k+1} - N_k < \frac{12}{\theta}. \tag{24}$$

La solution du problème à coefficients constants

$$\left.\begin{aligned} &av_{n-1} + bv_n + cv_{n+1} = f_n, \quad N_{k-1} < n < N_{k+1}, \\ &v_{N_{k-1}} = \varphi, \quad v_{N_{k+1}} = \psi, \end{aligned}\right\} \tag{25}$$

où

$$a = a_{N_k}, \quad b = b_{N_k}, \quad c = c_{N_k},$$

vérifie pour $n = N_k$, en vertu des inégalités

$$N_{k-1} + \frac{6}{\theta} < N_k < N_{k+1} - \frac{6}{\theta},$$

l'estimation (18) :

$$|v_{N_k}| \leqslant M_1 \max_m |f_m| + M_2' \max(|\varphi|, |\psi|),$$

avec

$$M_1 = \frac{128}{B\theta^3}, \quad M_2' = \frac{1}{5}.$$

Le problème

$$\left.\begin{aligned} &a_n u_{n-1} + b_n u_n + c_n u_{n+1} = f_n, \quad N_{k-1} < n < N_{k+1}, \\ &u_{N_{k-1}} = \varphi_k, \quad u_{N_{k+1}} = \psi_k \end{aligned}\right\} \tag{26}$$

peut être considéré comme perturbation du problème (25), l'écart entre les coefficients de (26) et ceux de (25) étant, en vertu de $N_{k+1} - N_{k-1} \leqslant 24/\theta$, au plus égal à ε. Nous pouvons utiliser l'inégalité (12) évaluant la solution du problème perturbé. Lorsque $n = N_k$, on a

$$|u_{N_k}| \leqslant 2M_1 \max_m |f_m| + \left(M_2' + \frac{1}{4}\right) \max(|u_{N_{k-1}}|, |u_{N_{k+1}}|) \leqslant$$
$$\leqslant 2M_1 \max_m |f_m| + \frac{1}{2} \max(|u_{N_{k-1}}|, |u_{N_{k+1}}|).$$

Par conséquent,

$$\max_{0<k<r} |u_{N_k}| \leqslant 2M_1 \max_m |f_m| + \frac{1}{2} \max\left[|\varphi|, |\psi|, \max_{0<k<r} |u_{N_k}|\right] \leqslant$$
$$\leqslant 2M_1 \max_m |f_m| + \frac{1}{2} \max_{0<k<r} |u_{N_k}| + \frac{1}{2} \max(|\varphi|, |\psi|).$$

D'où

$$\max_{0<k<r} |u_{N_k}| \leqslant 4M_1 \max_m |f_m| + \max(|\varphi|, |\psi|).$$

Trouvons maintenant, pour n arbitraire, ses deux limites N_{k-1} et N_{k+1} et servons-nous de l'estimation (23) :

$$|u_n| \leqslant 2M_1 \max_m |f_m| + 2M_2 \max(|u_{N_{k-1}}|, |u_{N_{k+1}}|) \leqslant$$
$$\leqslant 2M_1 \max_m |f_m| + 2M_2\left[4M_1 \max_m |f_m| + \max(|\varphi|, |\psi|)\right] \leqslant$$
$$\leqslant (2M_1 + 8M_1M_2) \max_m |f_m| + 2M_2 \max(|\varphi|, |\psi|). \tag{27}$$

En vertu de (23), l'estimation (27) obtenue pour $q - p > 24/\theta$ reste valable pour $q - p \leqslant 24/\theta$. Le problème (21) est résoluble pour des seconds membres quelconques puisqu'il découle de (27) que, lorsque $\varphi = \psi = f_m \equiv 0$, le problème n'a d'autre solution que 0.

Ainsi, nous avons démontré que dans les conditions de régularité (14) et sous les hypothèses (14'), la condition (15) constitue le critère de bon conditionnement du problème (1).

L'exemple ci-dessous montre qu'on ne saurait ignorer les conditions de régularité (14).

On vérifie aisément que le problème aux limites sous forme de différences

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= 0, \quad 0 < n < N, \\ u_0 = 0, \quad u_N &= 0, \end{aligned}\right\}$$

où $a_n \equiv 1$, $b_n = (-1)^n$, $c_n \equiv 1$ et $N = 6N_1$, admet pour tout N_1 entier naturel une solution non triviale

$$u_n = \begin{cases} \sin \dfrac{n\pi}{6} & \text{si } n \text{ est pair,} \\ -\cos \dfrac{n\pi}{6} & \text{si } n \text{ est impair.} \end{cases}$$

Ce problème n'est donc pas bien conditionné quoique

$$\frac{|b_n| - |a_n + c_n|}{|b_n| + |a_n| + |c_n|} = \frac{1}{3}, \quad |a_n| = |b_n| = |c_n| = 1,$$

i.e.

$$|q_1| < 1 - \frac{1}{6}, \quad |q_2^{-1}| < 1 - \frac{1}{6}.$$

3. Propriétés des problèmes bien conditionnés. Nous avons obtenu au §4 et au n° 2 du présent paragraphe certains résultats sur le bon conditionnement du problème (1). Enonçons-les de façon que nous puissions nous en servir en étudiant le balayage au §7.

Pour que le problème aux limites aux différences (1) soit bien conditionné, il suffit qu'il vérifie l'un des trois critères suivants :

premier critère :

$$|b_n| \geqslant |a_n| + |c_n| + \delta, \quad \delta > 0 ;$$

deuxième critère :

$$\frac{|b_n| - |a_n| - |c_n|}{|b_n| + |a_n| + |c_n|} \geqslant \theta > 0, \quad d_n = \max(|a_n|, |b_n|, |c_n|) \geqslant B > 0;$$

troisième critère :

$$\frac{|b_n| - |a_n + c_n|}{|b_n| + |a_n| + |c_n|} \geqslant \theta > 0, \quad M \geqslant d_n \geqslant B > 0,$$

les coefficients étant supposés réels et remplissant les conditions de régularité (14) :

$$|a_k - a_l| \leqslant D \left| \frac{k-l}{N} \right|^{\omega}, \quad |b_k - b_l| \leqslant D \left| \frac{k-l}{N} \right|^{\omega},$$

$$|c_k - c_l| \leqslant D \left| \frac{k-l}{N} \right|^{\omega}, \quad D > 0, \quad \omega > 0.$$

A condition de vérifier l'un quelconque des deux premiers critères, le problème (1) est résoluble pour $N \geqslant 2$ quels que soient les seconds membres, et il l'est pour tous les N suffisamment grands et pour des seconds membres arbitraires s'il satisfait au troisième. C'est également pour ces N qu'il existe

une solution de tous les problèmes aux limites « tronqués » de la forme (21).

La solution $\{u_n\}$ du problème originel et les solutions $\{u_n\}$ de tous les problèmes « tronqués » sont évaluées par

$$|u_n| \leqslant M \max\left(|\varphi|, |\psi|, \max_m |f_m|\right), \quad p \leqslant n \leqslant q,$$

avec M indépendant de N, p, q.

§7. Justification de la méthode du balayage pour des problèmes aux limites bien conditionnés

Nous sommes maintenant parfaitement armés pour aborder l'étude de la méthode du balayage décrite au §5. Soit à calculer la solution du problème aux limites sous forme de différences

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, \quad 0 < n < N, \\ u_0 = \varphi, \quad u_n &= \psi, \\ |a_n|, |b_n|, |c_n| &< M. \end{aligned}\right\} \tag{1}$$

Nous supposerons que ce problème et tous les problèmes qui s'en obtiennent par troncature

$$\left.\begin{aligned} a_n u_{n-1} + b_n u_n + c_n u_{n+1} &= f_n, \quad p < n < q, \\ u_p = \varphi, \quad u_q &= \psi, \end{aligned}\right\}$$

admettent des solutions $\{u_n\}$, les seconds membres étant quelconques ; de plus

$$|u_n| \leqslant M \max\left(|\varphi|, |\psi|, \max_m |f_m|\right). \tag{2}$$

Nous allons utiliser le fait qu'en vertu des estimations (4) et (5) du n° 1 §6, le problème aux différences à coefficients perturbés

$$\left.\begin{aligned} \tilde{a}_n \tilde{u}_{n-1} + \tilde{b}_n \tilde{u}_n + \tilde{c}_n \tilde{u}_{n+1} &= f_n, \quad 0 < n < N, \\ \tilde{u}_0 = \varphi, \quad \tilde{u}_N &= \psi, \\ |\tilde{a}_n - a_n|, |\tilde{b}_n - b_n|, |\tilde{c}_n - c_n| &\leqslant \varepsilon < \frac{1}{6M}, \end{aligned}\right\} \tag{3}$$

ainsi que tous les problèmes obtenus par troncature de (3), admet des solutions $\{\tilde{u}_n\}$ pour des seconds membres arbitraires ; de plus

$$|\tilde{u}_n| \leqslant 2M \max\left(|\varphi|, |\psi|, \max_m |f_m|\right). \tag{4}$$

1. Evaluation des coefficients de balayage. Nous montrerons qu'en calculant les coefficients de balayage nous n'aurons jamais à diviser par zéro, et nous obtiendrons des estimations de ces coefficients qui soient valables pour le problème initial (1) aussi bien que pour le problème perturbé (3).

Il suffit pour cela de nous limiter au problème perturbé puisque le problème de départ en est un cas particulier (pour $\varepsilon = 0$).

Examinons le système tronqué

$$\left.\begin{aligned} \tilde{a}_n\tilde{u}_{n-1}+\tilde{b}_n\tilde{u}_n+\tilde{c}_n\tilde{u}_{n+1} &= f_n, \quad 0 < n < l, \\ \tilde{u}_0 = \varphi, \quad \tilde{u}_l &= \psi, \end{aligned}\right\}$$

qui admet une solution, et trouvons $\tilde{u}_{l-1}$. Les formules de Cramer pour les solutions des systèmes d'équations algébriques linéaires entraînent que $\tilde{u}_{l-1}$ s'obtient comme

$$\tilde{u}_{l-1} = L\psi+\sum_{i=1}^{l-1} \Lambda_i f_i+\Lambda_0\varphi = L\tilde{u}_l+K, \tag{5}$$

avec L et Λ_i ne dépendant que de $\tilde{a}_n$, $\tilde{b}_n$, $\tilde{c}_n$. Vu l'estimation (4) qu'on a pour φ, ψ, $\{f_m\}$ arbitraires, il en résulte pour $\varphi = 0, f_i \equiv 0$, $\psi = 1$,

$$L = |\tilde{u}_{l-1}| \leqslant 2M$$

et

$$K = |\tilde{u}_{l-1}| \leqslant 2M \max\left(|\varphi|, \max_m |f_m|\right)$$

pour $\tilde{u}_l = \psi = 0$.

Il est commode d'indicer L et K par $l-{}^1/_2$ et d'écrire comme suit les relations et inégalités obtenues :

$$\left.\begin{aligned} \tilde{u}_{l-1} &= L_{l-1/2}\tilde{u}_l+K_{l-1/2}, \\ |L_{l-1/2}| \leqslant 2M, \quad |K_{l-1/2}| &\leqslant 2M \max\left(|\varphi|, \max_m |f_m|\right). \end{aligned}\right\} \tag{6}$$

En décrivant le balayage au §5, nous avons abouti à une relation analogue. Les formules de Cramer (5) montrent que $L_{l-1/2}$ est défini de façon unique en fonction des coefficients $\tilde{a}_n$, $\tilde{b}_n$, $\tilde{c}_n$ et que $K_{l-1/2}$ l'est en fonction de φ, f_n, $\tilde{a}_n$, $\tilde{b}_n$, $\tilde{c}_n$ $(0 < n < l)$. D'où la coïncidence de $L_{l-1/2}$, $K_{l-1/2}$ et des coefficients de balayage obtenus au §5 et pour lesquels nous y avons écrit les formules de récurrence

$$\left.\begin{aligned} L_{1/2} &= 0, \quad K_{1/2} = \varphi, \\ L_{l+1/2} &= \frac{-\tilde{c}_l}{\tilde{b}_l+\tilde{a}_l L_{l-1/2}}, \quad K_{l+1/2} = \frac{f_l-\tilde{a}_l K_{l-1/2}}{\tilde{b}_l+\tilde{a}_l L_{l-1/2}}. \end{aligned}\right\} \tag{7}$$

Il va de soi que la dernière affirmation n'est juste que si les formules de récurrence ont un sens, i.e. si aucun des dénominateurs y figurant ne s'annule.

Démontrons que les dénominateurs répondent à cette condition.

Supposons que les formules (7) permettent de calculer

$$\left.\begin{aligned} &L_{1/2}, L_{2/2}, \ldots, L_{l-1/2}, \\ &K_{1/2}, K_{2/2}, \ldots, K_{l-1/2} \end{aligned}\right\} l \geqslant 1$$

et testons-les en ce qui concerne $L_{l+1/2}$, $K_{l+1/2}$. Pour cela il suffit de montrer que

$$|\tilde{b}_l+\tilde{a}_l L_{l-1/2}| \geqslant \frac{1}{2M}. \tag{8}$$

Considérons le système

$$\left.\begin{aligned} \tilde{u}_0 &= 0, \\ \tilde{a}_i\tilde{u}_{i-1}+\tilde{b}_i\tilde{u}_i+\tilde{c}_i\tilde{u}_{i+1} &= 0, \quad i = 1, 2, \ldots, l-1, \\ \tilde{a}_l u_{l-1}+\tilde{b}_l\tilde{u}_l+\tilde{c}_l u_{l+1} &= 1, \\ \tilde{u}_{l+1} &= 0, \end{aligned}\right\} \tag{9}$$

dont on sait qu'il admet une solution. Les l premières équations (homogènes) impliquent $\tilde{u}_{l-1} = L_{l-1/2}\tilde{u}_l$, et la condition (4) entraîne $|\tilde{u}_l| \leqslant 2M$. Il s'ensuit de l'unique équation non homogène du système (9) que

$$(\tilde{a}_l L_{l-1/2}+\tilde{b}_l)\tilde{u}_l = 1.$$

Donc

$$\frac{1}{|\tilde{b}_l+\tilde{a}_l L_{l-1/2}|} \leqslant 2M,$$

ce qui démontre les estimations (8) et (6) et prouve en même temps que les formules de récurrence (7) ont un sens.

2. Influence des erreurs d'arrondi sur le résultat. Résolvons le problème (1) par la méthode du balayage. Lors des calculs effectifs, on commet à chaque pas des erreurs dues à l'arrondi. Le calcul réel se fait donc selon les formules

$$\left.\begin{aligned} L_{1/2} &= 0, \quad K_{1/2} = \varphi+\varkappa_{1/2}, \\ L_{l+1/2} &= \frac{-c_l}{a_l L_{l-1/2}+b_l}+\lambda_{l+1/2}, \qquad l = 1, 2, \ldots, N-1, \\ K_{l+1/2} &= \frac{f_l-a_l K_{l-1/2}}{a_l L_{l-1/2}+b_l}+\varkappa_{l+1/2}, \qquad l = 1, 2, \ldots, N-1, \\ u_N &= \psi+\nu_N, \\ u_l &= L_{l+1/2}u_{l+1}+K_{l+1/2}+\nu_l, \qquad l = N-1, N-2, \ldots, 1. \end{aligned}\right\} \tag{10}$$

Supposons que pour toutes les erreurs de calcul on a les estimations

$$|\varkappa_{l+1/2}| < \delta, \quad |\lambda_{l+1/2}| < \delta, \quad |\nu_l| < \delta$$

avec un δ suffisamment petit

$$\delta < \frac{1}{6M^2(2M+1)}.$$

Montrons que dans ce cas aucun des dénominateurs des formules (10) ne devient nul et apprécions l'influence des erreurs sur le résultat.

Notons

$$K_{1/2} = \tilde{K}_{1/2}; \quad K_{l+1/2}+\nu_l = \tilde{K}_{l+1/2}, \quad l > 0.$$

Il est évident que le résumé des formules (10) se met sous la forme

$$\left.\begin{aligned}
&L_{1/2} = 0, \quad \tilde{K}_{1/2} = \varphi + \varkappa_{1/2},\\
&L_{l+1/2} = -\frac{c_l - (a_l L_{l-1/2} + b_l)\lambda_{l+1/2}}{a_l L_{l-1/2} + b_l}, \quad l = 1, 2, \ldots, N-1,\\
&\tilde{K}_{l+1/2} = \frac{f_l - a_l(\tilde{K}_{l-1/2} - \nu_{l-1})}{a_l L_{l-1/2} + b_l} + \varkappa_{l+1/2} + \nu_l =\\
&= \frac{f_l + a_l \nu_{l-1} + (a_l L_{l-1/2} + b_l)(\varkappa_{l+1/2} + \nu_l) - a_l \tilde{K}_{l-1/2}}{a_l L_{l-1/2} + b_l}, \quad l = 1, 2, \ldots, N-1,\\
&u_N = \psi + \nu_N,\\
&u_l = L_{l+1/2} u_{l+1} + \tilde{K}_{l+1/2}, \quad l = N-1, N-2, \ldots, 1,
\end{aligned}\right\} \tag{10'}$$

et il est pris pour schéma de calcul pour le problème aux limites aux différences

$$\left.\begin{aligned}
\tilde{a}_n \tilde{u}_{n-1} + \tilde{b}_n \tilde{u}_n + \tilde{c}_n \tilde{u}_{n+1} = \tilde{f}_n, \quad 0 < n < N,\\
\tilde{u}_0 = \varphi, \quad \tilde{u}_N = \psi,
\end{aligned}\right\}$$

les seconds membres et les coefficients perturbés étant

$$\left.\begin{aligned}
&\tilde{\varphi} = \varphi + \varkappa_{1/2},\\
&\tilde{f}_l = f_l + a_l \nu_{l-1} + (a_l L_{l-1/2} + b_l)(\varkappa_{l+1/2} + \nu_l),\\
&\tilde{\psi} = \psi + \nu_N,\\
&\tilde{a}_l = a_l, \quad \tilde{b}_l = b_l, \quad \tilde{c}_l = c_l - (a_l L_{l-1/2} + b_l)\lambda_{l+1/2}.
\end{aligned}\right\} \tag{11}$$

Démontrons que

$$|\tilde{c}_l - c_l| \leqslant M(2M+1)\delta < \frac{1}{6M}. \tag{12}$$

On démontrera par récurrence sur l. Pour $l = 1$

$$|\tilde{c}_1 - c_1| = |(a_1 L_{1/2} + b_1)\lambda_{3/2}| = |(a_1 \cdot 0 + b_1)\lambda_{3/2}| \leqslant M\delta \leqslant M(2M+1)\delta < \frac{1}{6M}.$$

Supposons l'inégalité (12) démontrée pour $k = 1, 2, \ldots, l-1$. Pour calculer les coefficients $L_{1/2}, L_{3/2}, \ldots, L_{l-1/2}$ on utilisera seulement $\tilde{a}_l = a_l$, $\tilde{b}_i = b_l$ et $\tilde{c}_i$, $i = 1, 2, \ldots, l-1$. On peut donc affirmer, en vertu de (6), que $|L_{l-1/2}| \leqslant 2M$ et que, par conséquent,

$$|\tilde{c}_l - c_l| = |-(a_l L_{l-1/2} + b_l)\lambda_{l+1/2}| \leqslant (M \cdot 2M + M)\delta < \frac{1}{6M}.$$

Le raisonnement par récurrence est terminé.

Ainsi, nous avons montré que si $\delta \leqslant \frac{1}{6M^2(2M+1)}$, on a les inégalités

$$|\tilde{a}_n - a_n| = 0 < \frac{1}{6M}, \quad |\tilde{b}_n - b_n| = 0 < \frac{1}{6M}, \quad |\tilde{c}_n - c_n| < \frac{1}{6M},$$

et donc les estimations (6) et (8) :

$$\left.\begin{aligned} |L_{l-1/2}| &< 2M, \\ |\tilde{a}_l L_{l-1/2}+\tilde{b}_l| = |a_l L_{l-1/2}+b_l| &> \frac{1}{2M}. \end{aligned}\right\} \tag{13}$$

Le calcul par les formules (10) n'exige donc aucune division par zéro.

Les formules (11) donnant $\tilde{\varphi}, \tilde{f}_l, \tilde{\psi}$ et les estimations (13) entraînent maintenant les inégalités

$$|\tilde{\varphi}-\varphi| < \delta, \quad |\tilde{\psi}-\psi| < \delta,$$

$$|\tilde{f}_l-f_l| \leqslant M\delta+(M\cdot 2M+M)2\delta = M(4M+3)\delta.$$

En effectuant à chaque pas des calculs des erreurs au plus égales à δ, $\delta < \frac{1}{6M^2(2M+1)}$, nous résolvons donc le système à coefficients et seconds membres perturbés.

Ces perturbations sont au plus $M^*\delta$, où

$$M^* = \max\{2, (4M+3)M\}$$

dépend de M seul, les coefficients étant perturbés d'au plus $1/(6M)$.

De telles perturbations des coefficients et des seconds membres déterminent, comme le montre l'estimation (10) §6, des erreurs sur u_n qui sont au plus $M^{**}\delta$, M^{**} dépendant de M seul. (Si $M \approx N^r$, alors $M^* \approx N^{2r}$, $M^{**} \approx N^{3r}$, de sorte que l'erreur sur la solution sera $N^{3r}\delta$.)

Si M, et avec lui M^{**}, est indépendant de N et si on effectue à chaque pas de la méthode du balayage (le nombre de pas est proportionnel à N) des erreurs de l'ordre de δ, l'erreur résultante sera au plus const$\cdot\delta$.

Ainsi, l'influence sur le résultat de l'erreur commise à un pas ne croît pas avec N, et il en est également de l'influence totale des erreurs faites à tous les pas de calcul.

Cette propriété remarquable du balayage est à la base de sa popularité.

DEUXIÈME PARTIE

SCHÉMAS AUX DIFFÉRENCES POUR LES ÉQUATIONS DIFFÉRENTIELLES ORDINAIRES

La deuxième partie du présent ouvrage est consacrée à la construction et à l'étude de schémas aux différences pour les équations différentielles ordinaires. Nous introduirons les notions générales de convergence, d'approximation et de stabilité, notions qui sont essentielles dans la théorie des schémas aux différences. Le fait de se familiariser avec ces notions relativement aux équations différentielles ordinaires permettra, lorsqu'on aura à examiner des schémas aux différences pour les équations aux dérivées partielles, de se concentrer sur de nombreuses difficultés et particularités propres à cette classe vaste de problèmes.

CHAPITRE 4

EXEMPLES ÉLÉMENTAIRES DE SCHÉMAS AUX DIFFÉRENCES

Les schémas aux différences que nous introduirons dans ce chapitre poursuivent le seul but d'initier le lecteur aux notions théoriques fondamentales.

§8. Notions d'ordre de précision et d'approximation

1. Ordre de précision d'un schéma aux différences. Ce paragraphe traite de la convergence des solutions d'équations aux différences, lorsqu'on passe à un réseau plus fin, vers les solutions d'équations différentielles dont elles sont des approximations. Nous allons nous borner à deux schémas aux différences pour la résolution numérique du problème

$$\left.\begin{aligned} \frac{du}{dx}+Au = 0, \quad 0 \leqslant x \leqslant 1, \\ u(0) = b. \end{aligned}\right\} \tag{1}$$

Le premier schéma à considérer sera un schéma très simple utilisant l'équation aux différences

$$\frac{u(x+h)-u(x)}{h}+Au(x) = 0. \tag{2}$$

Partageons l'intervalle [0, 1] par des points distants de h. Pour des raisons de commodité, on choisit $h = 1/N$, N étant un entier. Numérotons les points de division de gauche à droite de façon que $x_n = nh$, $n = 0, 1, \ldots, N$. La valeur de u fournie par le schéma aux différences au point x_n sera notée u_n. Donnons-nous une valeur de départ et posons $u_0 = b$. L'équation aux différences (2) entraîne la relation

$$u_n = (1-Ah)u_{n-1},$$

qui permet de trouver la solution de (2) pour la condition initiale $u_0 = b$:

$$u_n = (1-Ah)^n b = (1-Ah)^{x_n/h} b. \tag{3}$$

Quant à la solution exacte du problème (1), elle est de la forme $u(x) = be^{-Ax}$. Au point x_n elle prend la valeur

$$u(x_n) = be^{-Ax_n}. \tag{4}$$

Evaluons maintenant l'erreur sur la solution approchée (3). Cette erreur est en x_n :

$$\delta(x_n) = [(1-Ah)^{x_n/h} - e^{-Ax_n}]b. \tag{5}$$

Comment décroît $\delta(x_n)$ avec l'augmentation du nombre de points de division ou, ce qui revient au même, avec la diminution du pas $h = 1/N$ du réseau aux différences ? Mettons $(1-Ah)^{x_n/h}$ sous la forme

$$(1-Ah)^{\frac{x_n}{h}} = e^{\frac{x_n}{h}\ln(1-Ah)} = e^{\frac{x_n}{h}\left[-Ah+\frac{A^2h^2}{2}+O(h^3)\right]} =$$

$$= e^{-Ax_n} e^{A^2\frac{h}{2}x_n} e^{O(h^2)} = e^{-Ax_n}\left[1+\frac{A^2hx_n}{2}+O(h^2)\right]\times$$

$$\times[1+O(h^2)] = e^{-Ax_n} + h\frac{A^2x_n}{2}e^{-Ax_n} + O(h^2).$$

L'égalité (3) se récrit donc

$$u_n = be^{-Ax_n} + hb\frac{A^2x_n}{2}e^{-Ax_n} + O(h^2) \tag{3'}$$

et

$$\delta(x_n) = hb\frac{A^2x_n}{2}e^{-Ax_n} + O(h^2) = O(h), \tag{6}$$

i.e. l'erreur (5) tend vers zéro avec h et elle est de l'ordre de la première puissance du pas.

Ceci autorise à dire que *l'ordre de précision du schéma aux différences est* 1 (ne pas confondre avec l'ordre d'une équation aux différences, notion que nous avons définie au § 1).

Résolvons à présent le problème (1) à l'aide de l'équation aux différences

$$\frac{u(x+h)-u(x-h)}{2h} + Au(x) = 0. \tag{7}$$

La résolution est plus ardue qu'elle ne paraît de prime abord. En effet, le

schéma considéré est une équation aux différences du second ordre, i.e. il faut que soient données deux conditions initiales $u(x_0) = u(0)$ et $u(x_1) = u(h)$, tandis que l'équation à intégrer (1) est du premier ordre et on ne définit dans ce cas que $u(0) = b$. Il est naturel de poser $u_0 = b$ également dans le schéma aux différences.

Mais comment définir u_1 ? Pour répondre à cette question utilisons la forme explicite de la solution de (7) (voir §3, formules (6)) :

$$u_n = u_0\left[\frac{q_2}{q_2-q_1}q_1^n - \frac{q_1}{q_2-q_1}q_2^n\right] + u_1\left[-\frac{1}{q_2-q_1}q_1^n + \frac{1}{q_2-q_1}q_2^n\right] = \\ = \frac{q_2u_0-u_1}{q_2-q_1}q_1^n - \frac{q_1u_0-u_1}{q_2-q_1}q_2^n, \tag{8}$$

où

$$\left.\begin{aligned} q_1 &= \sqrt{1+A^2h^2} - Ah = 1 - Ah + \frac{A^2h^2}{2} + O(h^4), \\ q_2 &= (-1)\left(1 + Ah + \frac{A^2h^2}{2}\right) + O(h^4). \end{aligned}\right\} \tag{9}$$

Les développements tayloriens (9) des racines de l'équation caractéristique permettent de donner des expressions approchées de q_1^n et q_2^n. Etablissons-les pour q_1^n :

$$q_1^n = q_1^{x_n/h} = \left[1 - Ah + \frac{A^2h^2}{2} + O(h^4)\right]^{x_n/h} = \\ = e^{\frac{x_n}{h}\ln\left[1 - Ah + \frac{A^2h^2}{2} + O(h^4)\right]}.$$

Puisque $\ln(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} + O(z^4)$, nous avons

$$\ln\left[1 - Ah + \frac{A^2h^2}{2} + O(h^4)\right] = -Ah + \frac{A^3h^3}{6} + O(h^4).$$

C'est pourquoi

$$q_1^n = e^{\frac{x_n}{h}\left[-Ah + \frac{A^3h^3}{6} + O(h^4)\right]} = e^{-Ax_n}\left[1 + h^2\frac{A^3x_n}{6}\right] + O(h^3). \tag{10}$$

En ce qui concerne q_2^n nous donnons le résultat tout prêt :

$$q_2^n = (-1)^n e^{Ax_n} + O(h^2). \tag{11}$$

Portons ces expressions approchées de q_1^n et q_2^n dans (8), il vient

$$u_n = \frac{q_2u_0-u_1}{q_2-q_1}q_1^n - \frac{q_1u_0-u_1}{q_2-q_1}q_2^n = \\ = \frac{q_2u_0-u_1}{q_2-q_1}\left[e^{-Ax_n} + h^2\frac{A^3x_n}{6}e^{-Ax_n} + \right. \\ \left. + O(h^3)\right] - \frac{q_1u_0-u_1}{q_2-q_1}(-1)^n[e^{Ax_n} + O(h^2)]. \tag{12}$$

Toutes nos déductions ultérieures se baseront sur cette formule.

Notons que si le coefficient $\frac{q_2u_0-u_1}{q_2-q_1}$ tend pour $h \to 0$ vers une limite finie b, le premier terme $\frac{q_2u_0-u_1}{q_2-q_1}\, q_1^n$ du second membre de (12) tend vers la solution cherchée du problème (1).

Comme

$$(-1)^n[e^{Ax_n}+O(h^2)] \xrightarrow[h\to 0]{} \begin{cases} e^{Ax_n}, & n \text{ étant pair,} \\ -e^{Ax_n}, & n \text{ étant impair,} \end{cases}$$

i.e. ne tend pas vers une limite déterminée, la convergence vers une limite pour $h \to 0$ du second terme

$$\frac{q_1u_0-u_1}{q_2-q_1}(-1)^n[e^{Ax_n}+O(h^2)] \tag{13}$$

a lieu dans la condition nécessaire que $\frac{q_1u_0-u_1}{q_2-q_1}$ tende vers zéro avec h.

Faisons le bilan.

Pour que la solution de l'équation aux différences

$$\frac{u(x+h)-u(x-h)}{2h}+Au(x) = 0$$

converge vers la solution $u = be^{-Ax}$ du problème aux limites (1), il faut que

$$\frac{q_1u_0-u_1}{q_2-q_1} \to 0, \quad \frac{q_2u_0-u_1}{q_2-q_1} \to b. \tag{14}$$

N'oubliez pas que nous avons convenu de poser $u_0 = b$. Les conditions (14) nous donnent une idée de façon dont on peut se donner u_1, à savoir il suffit que $u_1 \to u_0 = b$ pour $h \to 0$. En effet, $q_1 \to +1$, $q_2 \to -1$ pour $h \to 0$, et donc pour h tendant vers zéro

$$\frac{q_1u_0-u_1}{q_2-q_1} \to 0, \quad \frac{q_2u_0-u_1}{q_2-q_1} \to b.$$

2. Vitesse de convergence de la solution d'une équation aux différences. Etudions la vitesse de convergence pour divers choix concrets de $u_1 \approx u(h)$.

Pour définir $u(h)$ on utilise naturellement le développement taylorien de la solution de l'équation différentielle $u'+Au = 0$. Etant donné qu'en vertu de cette équation $u' = -Au$, récrivons la formule de Taylor

$$u(x_1) = u(0)-hAu(0)+O(h^2) = u(0)\,(1-Ah)+O(h^2).$$

Cette égalité a lieu pour la solution exacte de l'équation différentielle. S'il s'agit d'une solution approchée, on tronque ce développement au deuxième terme et on pose

$$u_1 = u_0(1-Ah).$$

En se limitant au premier terme, on pose

$$u_1 = u_0.$$

Dans le premier cas, la valeur initiale u_1 est entachée d'une erreur en h^2, et, dans le second, l'erreur est de l'ordre de h.

Etablissons la vitesse de convergence dans les deux cas.

Posons

$$u_0 = b, \quad u_1 = (1-Ah)b. \tag{15}$$

Alors (voir formules (9))

$$\frac{q_1u_0-u_1}{q_2-q_1} = \frac{[1-Ah+O(h^2)]b-(1-Ah)b}{-2+O(h^2)} = O(h^2),$$

$$\frac{q_2u_0-u_1}{q_2-q_1} = \frac{\left[-1-Ah-\frac{A^2h^2}{2}+O(h^4)\right]b-(1-Ah)b}{-2+O(h^4)} = b+O(h^2). \tag{16}$$

Compte tenu de l'égalité (12), nous obtenons sans peine le résultat recherché

$$u_n = be^{-Ax_n}+O(h^2), \tag{17}$$

qui s'énonce comme suit. Si la valeur initiale u_1 est donnée avec une précision en h^2, l'erreur sur la solution sera elle aussi de l'ordre de h^2, i.e. le schéma aux différences est *exact à l'ordre deux*.

On montre que même en se donnant $u_1 = be^{-Ax_1}$ (valeur exacte), la précision de la solution sera en h^2 au plus. Nous conseillons au lecteur de démontrer cette affirmation.

De même, on vérifie aisément qu'en se donnant un u_0 égal non pas à b, mais à toute quantité de la forme $b+O(h^2)$, la vitesse de convergence restera d'ordre deux.

Considérons maintenant le second cas et posons

$$u_1 = u_0 = b.$$

Alors

$$\frac{q_1u_0-u_1}{q_2-q_1} = \frac{[1-Ah+O(h^2)]b-b}{-2+O(h^2)} = \frac{1}{2}Ahb+O(h^2),$$

$$\frac{q_2u_0-u_1}{q_2-q_1} = \frac{-[1+Ah+O(h^2)]b-b}{-2+O(h^2)} = b+\frac{1}{2}Ahb+O(h^2)$$

et donc

$$u_n = \frac{q_2u_0-u_1}{q_2-q_1}[e^{-Ax_n}+O(h^2)]-\frac{q_1u_0-u_1}{q_2-q_1}(-1)^n[e^{Ax_n}+O(h^2)] =$$

$$= \left[b+\frac{1}{2}Ahb+O(h^2)\right][e^{-Ax_n}+O(h^2)]-$$

$$-(-1)^n\left[\frac{1}{2}Ahb+O(h^2)\right][e^{Ax_n}+O(h^2)] =$$

$$= be^{-Ax_n}+Ab\,\frac{e^{-Ax_n}-(-1)^ne^{Ax_n}}{2}h+O(h^2).$$

Ainsi, une erreur en h sur les données initiales entraîne une erreur du même ordre de grandeur sur la solution.

Résumons ce numéro. Contrairement au schéma

$$\frac{u(x+h)-u(x)}{h}+Au(x) = 0,$$

le schéma aux différences considéré

$$\frac{u(x+h)-u(x-h)}{2h}+Au(x) = 0$$

peut assurer une vitesse de convergence plus élevée, notamment une convergence avec un reste en h^2 et non pas en h, comme c'est le cas du premier schéma. Pour que le schéma soit exact à l'ordre deux, il faut, pour un u_0 donné exactement, choisir un u_1 tel que son écart à la valeur de la solution exacte de l'équation différentielle au point $x = x_0+h$ soit de l'ordre de h^2. On pourrait montrer qu'avec un u_0 entaché d'une erreur en h^2 l'ordre de convergence reste le même. Une précision de l'ordre de h^3 et plus des données initiales n'améliore pas l'exactitude de la solution.

En se donnant des données intiales affectées d'une erreur en h, on obtient une solution avec une erreur du même ordre de grandeur.

3. Ordre d'approximation. Voyons pourquoi le schéma

$$\frac{u(x+h)-u(x)}{h}+Au(x) = 0$$

s'avère moins précis que

$$\frac{u(x+h)-u(x-h)}{2h}+Au(x) = 0.$$

Ces schémas diffèrent par les expressions approchées

$$\frac{u(x+h)-u(x)}{h} \quad \text{et} \quad \frac{u(x+h)-u(x-h)}{2h}$$

de la dérivée du/dx au point x. Il est donc naturel de supposer que la première expression approchée est moins exacte. En effet, remplaçons $u(x+h)$ et $u(x-h)$ par leurs développements tayloriens

$$u(x+h) = u(x)+u'(x)h+u''(x)\frac{h^2}{2}+u'''(x)\frac{h^3}{6}+O(h^4),$$

$$u(x-h) = u(x)-u'(x)h+u''(x)\frac{h^2}{2}-u'''(x)\frac{h^3}{6}+O(h^4).$$

Compte tenu de ces développements, on a

$$\frac{u(x+h)-u(x)}{h} = u'(x)+u''(x)\frac{h}{2}+O(h^2),$$

$$\frac{u(x+h)-u(x-h)}{2h} = u'(x)+u'''(x)\frac{h^2}{6}+O(h^4),$$

i.e. la première approximation est exacte à l'ordre un et la seconde à l'ordre deux.

Les exemples examinés autorisent à supposer que l'ordre de convergence

des solutions des équations aux différences peut être ramené à l'ordre d'approximation des dérivées de l'équation différentielle.

Or, énoncée d'une façon si générale, cette hypothèse est fausse. Sa validité exige que les schémas aux différences soient assujettis à une condition essentielle, à savoir à être *stables*. La nécessité de cette contrainte ressortira de l'exemple considéré au paragraphe suivant.

§9. Un schéma aux différences instable

1. Procédés d'approximation d'une dérivée. Revenons aux schémas aux différences pour l'intégration approchée de l'équation différentielle simple $u'+Au = 0$. Pour former un schéma aux différences approchant cette équation il suffit, nous l'avons vu, de remplacer la dérivée u' par un rapport aux différences approximant. Ainsi, nous avons considéré des schémas pour lesquels la dérivée u' a été remplacée par

$$\frac{u(x+h)-u(x)}{h} \quad \text{ou} \quad \frac{u(x+h)-u(x-h)}{2h}.$$

Il est évident de même que toute expression de la forme

$$\mu\,\frac{u(x+h)-u(x-h)}{2h}+(1-\mu)\,\frac{u(x+h)-u(x)}{h}$$

approche $u'(x)$. En effet, substituons-y les développements tayloriens de $u(x+h)$ et $u(x-h)$:

$$u(x+h) = u(x)+u'(x)h+O(h^2),$$
$$u(x-h) = u(x)-u'(x)h+O(h^2).$$

Il vient

$$\mu\,\frac{u(x+h)-u(x-h)}{2h}+(1-\mu)\,\frac{u(x+h)-u(x)}{h} =$$
$$= \mu\,\frac{[u(x)+u'(x)h+O(h^2)]-[u(x)-u'(x)h+O(h^2)]}{2h}+$$
$$+(1-\mu)\,\frac{[u(x)+u'(x)h+O(h^2)]-u(x)}{h} = u'(x)+O(h).$$

Avec cette façon d'approcher une dérivée, on obtient toute une famille de schémas aux différences dépendant d'un paramètre numérique.

Ce seront des schémas du type

$$\mu\,\frac{u(x+h)-u(x-h)}{2h}+(1-\mu)\,\frac{u(x+h)-u(x)}{h}+Au(x) = 0. \qquad (1)$$

A chaque valeur du paramètre μ correspond un schéma déterminé. Au §8 nous avons étudié les schémas répondant à $\mu = 0$ et $\mu = 1$.

2. Exemple de schéma aux différences instable. Soit un autre schéma de ce type qui s'obtient à partir de (1) lorsque $\mu = 4$:

$$4\,\frac{u(x+h)-u(x-h)}{2h}-3\,\frac{u(x+h)-u(x)}{h}+Au(x) = 0. \qquad (2)$$

Ce schéma se récrit comme suit :

$$-2u(x-h)+(3+Ah)u(x)-u(x+h) = 0. \tag{2'}$$

Comme dans les exemples précédents, la solution sera obtenue sur l'intervalle [0, 1] partagé par les points du réseau aux différences en N intervalles partiels égaux de longueur $h = 1/N$. La coordonnée x_n d'un point du réseau est définie par $x_n = nh = n/N$.

La solution de l'équation aux différences s'explicite en

$$u_n = u_0\left[\frac{q_2}{q_2-q_1}q_1^n-\frac{q_1}{q_2-q_1}q_2^n\right]+u_1\left[-\frac{1}{q_2-q_1}q_1^n+\frac{1}{q_2-q_1}q_2^n\right], \tag{3}$$

q_1 et q_2 étant les racines de l'équation caractéristique

$$-2+(3+Ah)q-q^2 = 0.$$

Calculons ces racines. Nous avons

$$\left.\begin{aligned} q_1 &= \frac{3+Ah-\sqrt{1+6Ah+A^2h^2}}{2} = 1-Ah+2A^2h^2+O(h^3),\\ q_2 &= \frac{3+Ah+\sqrt{1+6Ah+A^2h^2}}{2} = 2(1+Ah)+O(h^2). \end{aligned}\right\} \tag{4}$$

Nous allons également utiliser les expressions approchées de q_1^n, q_2^n :

$$\left.\begin{aligned} q_1^n &= [1-Ah+O(h^2)]^n = [1-Ah+O(h^2)]^{x_n/h} =\\ &\qquad = e^{-Ax_n}+O(h),\\ q_2^n &= [2(1+Ah)+O(h^2)]^n = [2(1+Ah)+O(h^2)]^{x_n/h} =\\ &\qquad = 2^{x_n/h}[e^{Ax_n}+O(h)]. \end{aligned}\right\} \tag{5}$$

Portons les égalités (5) dans la formule (3), il vient

$$u_n = \frac{q_2u_0-u_1}{q_2-q_1}[e^{-Ax_n}+O(h)]+\frac{q_1u_0-u_1}{q_1-q_2}[e^{Ax_n}+O(h)]2^{x_n/h}. \tag{6}$$

Avant de mettre en évidence la limite vers laquelle tend u_n pour $h \to 0$, nous devons indiquer comment se définissent les valeurs initiales u_0 et u_1 de la solution de l'équation aux différences.

Imitons §8 et cherchons la solution qui vérifie la condition $u(0) = b$. Prenons pour données initiales aux différences $u_0 = b$ et $u_1 = b(1-Ah)$, portons-les dans la formule (6) et simplifions séparément chaque terme. Nous avons alors pour deux premiers termes à droite :

$$\frac{q_2u_0-u_1}{q_2-q_1}[e^{-Ax_n}+O(h)] = \frac{[2+O(h)]b-(1-Ah)b}{[2+O(h)]-[1-O(h)]}\times$$
$$\times[e^{-Ax_n}+O(h)] = be^{-Ax_n}+O(h),$$

$$\frac{q_1u_0-u_1}{q_2-q_1}[e^{Ax_n}+O(h)]2^{x_n/h} = \frac{[1-Ah+2A^2h^2+O(h^3)]b-b(1-Ah)}{[1+O(h)]-[2+O(h)]}\times$$
$$\times[e^{Ax_n}+O(h)]2^{x_n/h} = -2A^2h^2b[e^{Ax_n}+O(h)]2^{x_n/h}$$

et donc

$$u_n = [be^{-Ax_n}+O(h)]+[-2A^2be^{Ax_n}+O(h)]h^2 2^{x_n/h}.$$

Le premier terme de cette formule tend, lorsque $h \to 0$, $x_n = x = \text{const}$, vers be^{-Ax}, i.e. vers la solution cherchée. Pour que ce soit également la limite de toute l'expression de u_n, il faut donc que le second terme converge vers zéro. Or, pour $h \to 0$, celui-ci tend vers l'infini. En effet, $-2A^2be^{Ax_n}+$ $+O(h)$ a pour limite $-2A^2be^{Ax}$ qui est fini, $\neq 0$, et $h^2 2^{x_n/h}$ tend vers l'infini plus vite que toute puissance positive de $1/h$.

Nous avons montré que le schéma aux différences approximant une équation différentielle peut admettre une solution qui ne converge pas pour $h \to 0$ vers la solution de l'équation différentielle. On pourrait penser que la cause en est un choix insuffisamment précis de u_1. Nous allons cependant montrer qu'il n'y a pas de convergence même si u_1 est choisi exactement égal à la solution de l'équation différentielle pour $x_1 = x_0+h$, i.e. si l'on pose $u_1 = u_0e^{-Ah} = be^{-Ah}$. Commençons notre démonstration en simplifiant les expressions figurant dans (6) :

$$\frac{q_2u_0-u_1}{q_2-q_1} = \frac{[2+O(h)]b-bc^{-Ah}}{[2+O(h)]-[1+O(h)]} = b+O(h),$$

$$\frac{q_1u_0-u_1}{q_1-q_2} = \frac{[1-Ah+2A^2h^2+O(h^3)]b-be^{-Ah}}{[1+O(h)]-[2+O(h)]} =$$

$$= -\frac{3}{2}A^2h^2[b+O(h)].$$

Portons ces expressions dans (6), il vient

$$u_n = [be^{-Ax_n}+O(h)]-\left[\frac{3}{2}A^2be^{Ax_n}+O(h)\right]h^2 2^{x_n/h}. \tag{7}$$

Le second terme à droite tend de nouveau vers l'infini et le premier reste borné. D'où la conclusion suivante : toute la solution de l'équation aux différences tend vers l'infini.

La cause de cette absence de convergence pour $h \to 0$ du schéma aux différences (2) est qu'il admet des solutions croissant rapidement avec la diminution du pas h même pour des données initiales définies de façon raisonnable.

De tels schémas aux différences sont dits *instables*. On conçoit qu'on ne saurait les utiliser pour la résolution numérique des équations différentielles.

CHAPITRE 5

CONVERGENCE DES SOLUTIONS DES ÉQUATIONS AUX DIFFÉRENCES, CONSÉQUENCE DE L'APPROXIMATION ET DE LA STABILITÉ

Au chapitre 4 nous avons établi, des exemples aidant, la notion d'approximation d'un problème différentiel par un problème aux différences et celle de convergence qui fait que la solution du problème différentiel se prête à un calcul approché à l'aide d'un schéma aux différences. Nous avons appris ce qu'est l'instabilité qui peut être à l'origine de la divergence et de l'inefficacité numérique de ce schéma. En analysant le comportement des solutions dans ces exemples élémentaires préliminaires dont la seule ambition a été de faire connaître dans leurs grandes lignes les premières notions, nous avons écrit les solutions sous forme de formules. Cette écriture n'a été possible que grâce à une sélection spéciale d'exemples.

Nous nous proposons de donner dans le présent chapitre les définitions rigoureuses de la convergence, de l'approximation et de la stabilité et de montrer que la démonstration de la convergence peut ne pas s'appuyer sur l'analyse des formules donnant les solutions. En effet, elle peut se faire en deux étapes : la vérification de l'approximation du problème différentiel par celui aux différences et la vérification de la stabilité de ce dernier.

§10. Convergence d'un schéma aux différences

1. Réseau et fonction définie sur un réseau. Soit posé, sur un segment D, un problème aux limites différentiel. Cela veut dire qu'on se donne une équation différentielle (ou un système différentiel) que vérifie une solution u sur D et des conditions supplémentaires pour u à l'une ou aux deux extrémités de D. Le problème aux limites différentiel s'écrira sous forme d'égalité symbolique

$$Lu = f, \tag{1}$$

avec L un opérateur différentiel donné et f un second membre donné. Ainsi, pour mettre le problème

$$\left.\begin{aligned} \frac{du}{dx} + \frac{x}{1+u^2} &= \cos x, \quad 0 \leqslant x \leqslant 1, \\ u(0) &= 3 \end{aligned}\right\} \tag{2}$$

sous forme (1), il suffit de poser

$$Lu \equiv \begin{cases} \dfrac{du}{dx} + \dfrac{x}{1+u^2}, & 0 \leqslant x \leqslant 1, \\ u(0), \end{cases}$$

$$f \equiv \begin{cases} \cos x, & 0 \leqslant x \leqslant 1, \\ 3. \end{cases}$$

Le problème

$$\left.\begin{aligned} \frac{d^2u}{dx^2} - (1+x^2)u &= \sqrt{x}, \quad 0 \leqslant x \leqslant 1, \\ u(0) &= 2, \\ \frac{du(0)}{dx} &= 1 \end{aligned}\right\} \tag{3}$$

prend la forme (1) à condition de poser

$$Lu \equiv \begin{cases} \dfrac{d^2u}{dx^2} - (1+x^2)u, & 0 \leqslant x \leqslant 1, \\ u(0), \\ \dfrac{du(0)}{dx}, \end{cases}$$

$$f \equiv \begin{cases} \sqrt{x}, & 0 \leqslant x \leqslant 1, \\ 2, \\ 1. \end{cases}$$

Pour pouvoir donner la forme (1) au problème

$$\left.\begin{aligned} \frac{d^2u}{dx^2} - (1+x^2)u &= \sqrt{x+1}, \quad 0 \leqslant x \leqslant 1, \\ u(0) &= 2, \\ u(1) &= 1 \end{aligned}\right\} \tag{4}$$

avec les conditions aux deux extrémités du segment $0 \leqslant x \leqslant 1$, il faut poser

$$Lu \equiv \begin{cases} \dfrac{d^2u}{dx^2} - (1+x^2)u, & 0 \leqslant x \leqslant 1, \\ u(0), \\ u(1), \end{cases}$$

$$f \equiv \begin{cases} \sqrt{x+1}, & 0 \leqslant x \leqslant 1, \\ 2, \\ 1. \end{cases}$$

Le problème aux limites pour le système d'équations différentielles

$$\left.\begin{aligned} \frac{dv}{dx}+xvw &= x^2-3x+1, & 0 \leqslant x \leqslant 1,\\ \frac{dw}{dx}+\frac{1}{1+x^2}(v+w) &= \cos^2 x, & 0 \leqslant x \leqslant 1,\\ v(0) &= 1,\\ w(0) &= -3 \end{aligned}\right\} \tag{5}$$

s'écrira comme (1) si on assimile u à la fonction vectorielle $u = \begin{pmatrix} v \\ w \end{pmatrix}$ et si on pose

$$Lu \equiv \begin{cases} \dfrac{dv}{dx}+xvw, & 0 \leqslant x \leqslant 1,\\ \dfrac{dw}{dx}+\dfrac{1}{1+x^2}(v+w), & 0 \leqslant x \leqslant 1,\\ v(0),\\ w(0), \end{cases}$$

$$f \equiv \begin{cases} x^2-3x+1, & 0 \leqslant x \leqslant 1,\\ \cos^2 x, & 0 \leqslant x \leqslant 1,\\ 1,\\ -3. \end{cases}$$

Pour fixer les idées, nous avons considéré tous les problèmes précédents sur un même segment $0 \leqslant x \leqslant 1$.

Supposons que le problème (1) admet sur $0 \leqslant x \leqslant 1$ une solution $u(x)$. Le calcul de cette solution par la méthode des différences finies ou méthode des réseaux exige qu'on choisisse sur D un nombre fini de points dont l'ensemble s'appelle *réseau* et se note D_h, et qu'on cherche ensuite non pas la solution $u(x)$ du problème (1), mais la table $[u]_h$ de valeurs de $u(x)$ aux points de D_h. Le réseau D_h est supposé dépendant d'un paramètre $h > 0$ qui peut prendre des valeurs positives aussi petites qu'on le veut. Avec le pas h tendant vers zéro, le réseau devient toujours plus fin. Par exemple, on peut poser $h = 1/N$, N étant un entier naturel, et prendre pour D_h l'ensemble des points $x_0 = 0$, $x_1 = h$, $x_2 = 2h$, ..., $x_N = 1$. La fonction discrète cherchée $[u]_h$ prend alors au point $x_n = nh$ de D_h la valeur $u(nh)$ que nous noterons pour abréger u_n.

Dans le cas du problème (2), la table de valeurs de $[u]_h$ se calcule de façon approchée en recourant, par exemple, au système

$$\left.\begin{aligned} \frac{u_{n+1}-u_n}{h}+\frac{x_n}{1+u_n^2} &= \cos x_n, \quad n = 0, 1, \ldots, N-1,\\ u_0 &= 3, \end{aligned}\right\} \tag{6}$$

obtenu en remplaçant la dérivée du/dx aux points du réseau par un rapport aux différences selon la formule approchée

$$\frac{du}{dx} \approx \frac{u(x+h)-u(x)}{h}.$$

La solution $u^{(h)} = (u_0^{(h)}, u_1^{(h)}, \ldots, u_N^{(h)})$ du système (6) est définie sur le même réseau D_h que la fonction cherchée $[u]_h$. Ses valeurs $u_1^{(h)}, u_2^{(h)}, \ldots, u_N^{(h)}$ aux points $x_1, x_2, \ldots, x_N$ s'obtiennent successivement à partir de (6) pour $n = 0, 1, \ldots, N-1$. Pour être plus bref, nous omettons l'indice h des $u_n^{(h)}$ de (6). En règle générale, nous agirons de même dans des cas analogues.

On recherche la fonction discrète $u^{(h)}$ du problème (4), qui coïncide approximativement avec la table cherchée de valeurs de la solution $[u]_h$, en utilisant le schéma aux différences

$$\left.\begin{aligned} \frac{u_{n+1}-2u_n+u_{n-1}}{h^2} - (1+x_n^2)u_n &= \sqrt{x_n+1}, \quad n = 1, 2, \ldots, N-1, \\ u_0 &= 2, \\ u_N &= 1. \end{aligned}\right\} \tag{7}$$

Ce schéma s'obtient en substituant aux points du réseau à la dérivée d^2u/dx^2 un rapport aux différences par la formule approchée

$$\frac{d^2u}{dx^2} \approx \frac{u(x+h)-2u(x)+u(x-h)}{h^2}. \tag{8}$$

La solution $u^{(h)}$ du problème (7) se calcule par l'algorithme de balayage décrit au §5.

Et voici le schéma aux différences susceptible de donner la solution du problème (5) :

$$\left.\begin{aligned} &\frac{v_{n+1}-v_n}{h} + x_n v_n w_n = x_n^2 - 3x_n + 1, \\ &\frac{w_{n+1}-w_n}{h} + \frac{1}{1+x_n^2}(v_n+w_n) = \cos^2 x_n, \quad n = 0, 1, \ldots, N-1, \\ &v_0 = 1, \quad w_0 = -3. \end{aligned}\right\} \tag{9}$$

Ici $u_0^{(h)} = \begin{pmatrix} v_0^{(h)} \\ w_0^{(h)} \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \end{pmatrix}$ est défini. Lorsque $n = 0$, moyennant (9) on trouve $u_1^{(h)} = \begin{pmatrix} v_1^{(h)} \\ w_1^{(h)} \end{pmatrix}$. Connaissant $u_k^{(h)} = \begin{pmatrix} v_k^{(h)} \\ w_k^{(h)} \end{pmatrix}$, $k = 0, 1, \ldots, n$, on peut en général, pour $k = n$, calculer $u_{n+1}^{(h)} = \begin{pmatrix} v_{n+1}^{(h)} \\ w_{n+1}^{(h)} \end{pmatrix}$.

Dans les exemples considérés, les points du réseau D_h sont distants de h. On conçoit que $N+1$ points de D_h, $h \equiv 1/N$, sur $[0, 1]$ pourraient être non pas équidistants, mais tels que $x_0 = 0$, $x_1 = x_0+h_0$, $x_2 = x_1+h_1$, ..., $x_{n+1} = x_n+h_n$, $x_N = 1$, avec h_n, $n = 0, 1, \ldots, N-1$, des nombres inégaux mais tels que $\max_n h_n \to 0$ pour $h = 1/N \to 0$. En choisissant de façon con-

venable les nœuds de D_h on est amené à la table cherchée $[u]_h$ de la solution $u(x)$ qui est plus détaillée pour N (ou $h = 1/N$) fixe donné dans les secteurs de variation particulièrement rapide de $u(x)$. On est parfois renseigné d'avance sur de tels secteurs soit par des considérations physiques, soit par des calculs préliminaires grossiers. L'information sur la vitesse de variation de $u(x)$ apparaît également à mesure qu'on calcule successivement $u_1^{(h)}, u_2^{(h)}, \ldots, u_n^{(h)}$ et peut être prise en considération lors du choix du nœud suivant x_{n+1}.

Nous estimons que les exemples ci-dessus suffisent pour comprendre ce qu'est un réseau et une fonction discrète (ou vectorielle) cherchée (la table de valeurs de la solution $[u]_h$). Notons de plus qu'il n'est nullement nécessaire de prendre pour la table cherchée une fonction coïncidant avec la solution u aux points du réseau. Il existe d'autres procédés pour établir la correspondance entre une fonction et sa table. On peut prendre, par exemple, pour la table de $u(x)$, $0 < x < 1$, une fonction discrète $[u]_h$ définie aux points $x = \frac{h}{2}, \frac{3}{2}h, \ldots, 1-\frac{h}{2}$ par l'égalité

$$[u]_h = \frac{1}{h} \int_{x-h/2}^{x+h/2} u(\xi)\, d\xi.$$

Cela est commode si $u(x)$ n'est pas continue mais s'il existe une intégrale par rapport à $u(x)$ sur tout segment. Ce cas peut se présenter, par exemple, pour des solutions généralisées — discontinues —, s'il existe

$$\int_0^1 u^2(x)\, dx.$$

Dans la suite, nous supposerons (à moins que le contraire ne soit spécifié) que u est une fonction continue et $[u]_h$ une fonction coïncidant avec u aux points du réseau.

Si c'est la fonction $[u]_h$ qu'on demande de calculer, c'est qu'avec le passage à un réseau plus fin, i.e. pour $h \to 0$, cette table de u cherché se remplit progressivement et en donne une représentation toujours plus parfaite. On pourrait, avec une précision augmentant pour $h \to 0$, définir par interpolation la solution u partout dans le demaine D. Il est clair que la précision dépend, pour un nombre fixe de nœuds de D_h, de l'information complémentaire sur la solution (du type d'estimations de ses dérivées), ainsi que de la position de ces nœuds.

La détermination d'une fonction connaissant sa table constitue l'objet de la théorie de l'interpolation. Aussi nous n'allons pas approfondir le problème analogue pour la fonction u et sa table $[u]_h$ en nous consacrant exclusivement au calcul de celle-ci. Convenons donc que le problème (1) est résolu exactement si on a trouvé la fonction $[u]_h$. Il est cependant impossible de connaître $[u]_h$ exactement. C'est pourquoi nous chercherons au lieu de $[u]_h$ une autre fonction discrète $u^{(h)}$ qui « converge » vers $[u]_h$ en resserrant le pas du réseau.

2. Schémas aux différences convergents. Avant de passer à la construction et à l'étude de schémas aux différences convergents, qui nous occuperont tout au long de ce chapitre, voyons ce que signifie exactement la condition de convergence $u^{(h)} \to [u]_h$ que nous imposerons aux schémas aux différences. A cette fin considérons l'espace vectoriel normé de fonctions définies sur le réseau D_h. *La norme* $\|u_h\|_{U_h}$ *d'une fonction discrète* $u_h \in U_h$ est un nombre non négatif qui mesure l'écart entre u_h et la valeur identiquement nulle. Rappelons qu'un espace vectoriel R est dit *normé* si on a fait correspondre à tout élément x de cet espace un nombre non négatif $\|\mathbf{x}\|$, les trois axiomes de norme étant vérifiés :

1° $\|\mathbf{x}\| \geqslant 0, \quad \mathbf{x} \in R$;

2° $\|\lambda\mathbf{x}\| = |\lambda| \cdot \|\mathbf{x}\|$, où $\mathbf{x} \in R$, λ étant arbitraire ;

3° $\|\mathbf{x}+\mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$, où $\mathbf{x}, \mathbf{y} \in R$.

Une norme se définit de façons différentes. On prend, par exemple, pour norme d'une fonction la borne supérieure du module de ses valeurs aux points du réseau en posant

$$\|u_h\|_{U_h} = \sup_n |u_h(x_n)|. \tag{10}$$

Si $u^{(h)}$ est un couple de fonctions comme dans (9), la norme analogue à (10) est la borne supérieure des modules des deux fonctions sur les réseaux respectifs.

Dans le cas de $u^{(h)}$ formée de fonctions définies sur le réseau $x = 0, h, 2h, \ldots, 1$, on utilise souvent une norme définie par l'égalité

$$\|u^{(h)}\|_{U_h} = \left(h \sum_{n=0}^{N} |u_n|^2\right)^{1/2}.$$

Elle est analogue à

$$\|u(x)\| = \left(\int_0^1 |u(x)|^2\, dx\right)^{1/2}$$

dont sont munies les fonctions $u(x)$ de carré intégrable sur $0 \leqslant x \leqslant 1$.

Sauf mention du contraire, nous utiliserons partout la norme (10).

En introduisant l'espace normé U_h on donne par là même un sens à la notion d'écart de deux fonctions. Soient $a^{(h)}$ et $b^{(h)}$ deux fonctions quelconques de U_h définies sur un réseau. On *mesure l'écart* entre $a^{(h)}$ et $b^{(h)}$ par la norme de leur différence, i.e. le nombre

$$\|a^{(h)} - b^{(h)}\|_{U_h}.$$

Il est maintenant possible de définir rigoureusement un schéma aux différences convergent.

Supposons que le calcul approché de la solution du problème aux limites différentiel (1), i.e. le calcul approché de la fonction $[u]_h$ à l'aide de l'égalité (1), se fait moyennant un système d'équations que nous écrirons, comme (1),

sous forme d'égalité symbolique

$$L_h u^{(h)} = f^{(h)}. \tag{11}$$

Illustrons-le sur l'exemple des schémas (6), (7), (9) pour les problèmes (2), (4), (5) respectivement.

On écrit (6) sous forme (11) en posant

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{n+1}-u_n}{h} + \dfrac{nh}{1+u_n^2}, & n = 0, 1, \ldots, N-1, \\ u_0, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} \cos nh, & n = 0, 1, \ldots, N-1. \\ 3. \end{cases}$$

Le schéma (7) se met sous la forme (11) en prenant

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{n+1}-2u_n+u_{n-1}}{h^2} + [1-(nh)^2]u_n, & n = 1, 2, \ldots, N-1, \\ u_0, \\ u_N, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} \sqrt{1+nh}, & n = 1, 2, \ldots, N-1, \\ 2, \\ 1, \end{cases}$$

et le schéma (9) en prenant

$$L_h u^{(h)} \equiv L_h \begin{pmatrix} v^{(h)} \\ w^{(h)} \end{pmatrix} = \begin{cases} \dfrac{v_{n+1}-v_n}{h} + (nh) v_n w_n, & n = 0, 1, \ldots, N-1, \\ \dfrac{w_{n+1}-w_n}{h} + \dfrac{1}{1+(nh)^2}(v_n+w_n), & n = 0, 1, \ldots, N-1, \\ v_0, \\ w_0, \end{cases}$$

$$f^{(h)} = \begin{cases} (nh)^2 - 3nh + 1, & n = 0, 1, \ldots, N-1, \\ \cos^2 nh, & n = 0, 1, \ldots, N-1, \\ 1, \\ -3. \end{cases}$$

Le système (11) dépend donc de h et doit être écrit pour tous les h pour lesquels on considère le réseau D_h et la fonction discrète $[u]_h$. Ainsi, le problème aux limites aux différences (11) ne constitue pas un seul système : c'est une famille de systèmes dépendant du paramètre h.

Nous supposerons que pour tout h suffisamment petit il existe une solution $u^{(h)}$ du problème (11) appartenant à l'espace U_h.

Nous dirons que *la solution $u^{(h)}$ du problème aux différences* (11) *converge*, avec un réseau de plus en plus fin, *vers la solution u du problème différentiel* (1) si

$$\|[u]_h - u^{(h)}\|_{U_h} \to 0 \quad \text{pour} \quad h \to 0. \tag{12}$$

Si on a de plus l'inégalité

$$\|[u]_h - u^{(h)}\|_{U_h} \leqslant ch^k, \tag{13}$$

où $c > 0$ et $k > 0$ sont des constantes indépendantes de h, nous dirons que *la convergence est de l'ordre de h^k* ou que *le schéma aux différences est exact à l'ordre k.*

Au §8 nous avons examiné deux schémas aux différences pour le problème

$$\left.\begin{aligned} \frac{du}{dx} + Au = 0, \quad 0 \leqslant x \leqslant 1, \\ u(0) = b, \end{aligned}\right\}$$

et évalué la différence $\delta(x) = u(x_k) - u_k^{(h)}$ entre la solution exacte et celle approchée. Ces estimations équivalent à dire que la convergence du premier schéma est de l'ordre de h et celle du second de l'ordre de h^2.

La propriété de convergence est la première condition imposée au schéma aux différences (11) pour qu'il permette de résoudre numériquement le problème aux limites différentiel (1). Si cette propriété a lieu, on calcule moyennant le schéma (11) la solution u avec n'importe quelle précision donnée à l'avance à condition de choisir un h suffisamment petit. Nous avons donc formulé rigoureusement la notion de convergence et abordé cette question majeure qu'est la construction du schéma convergent (11) destiné à la résolution du problème différentiel (1). Les exemples ci-dessus complètent ceux du chapitre premier et donnent une idée du procédé de construction le plus rudimentaire : on choisit un réseau et on remplace les dérivées par des rapports aux différences. Or, pour un même problème aux limites différentiel, on obtient, nous l'avons vu, divers schémas aux différences (11) à force de choisir de façon différente le réseau D_h et de remplacer différemment les dérivées par des rapports aux différences approximants. En étudiant une équation différentielle ordinaire simple (§6), nous avons constaté qu'un schéma aux différences peut s'avérer impraticable.

3. Vérification de la convergence d'un schéma aux différences. Laissons pour le moment de côté la construction des schémas aux différences et posons le problème en des termes un peu différents. Supposons construit, par le jeu de certaines considérations, un schéma aux différences $L_h u^{(h)} = f^{(h)}$ tel qu'on puisse espérer la convergence

$$\|[u]_h - u^{(h)}\|_{U_h} \to 0 \quad \text{pour } h \to 0.$$

Comment s'assurer si la convergence a effectivement lieu ?

Admettons que le problème aux différences (11) possède une solution unique $u^{(h)} \in U_h$. Si, par substitution de la fonction discrète $[u]_h \in U_h$ à $u^{(h)}$ l'égalité (11) se trouvait rigoureusement vérifiée, alors, la solution étant

unique, on aurait $[u]_h = u^{(h)}$, égalité idéale du point de vue convergence. Autrement dit, la solution $u^{(h)}$ du problème aux différences $L_h u^{(h)} = f^{(h)}$ aurait coïncidé avec la fonction cherchée $[u]_h$ que nous avons convenu d'estimer être la solution exacte.

Or, il est en général impossible de choisir le système (11) de façon que $[u]_h$ le vérifie exactement. En portant $[u]_h$ dans l'équation (11) on voit surgir un résidu

$$L_h[u]_h = f^{(h)} + \delta f^{(h)}. \tag{14}$$

Si ce résidu $\delta f^{(h)}$ « tend vers zéro » avec h de sorte que $[u]_h$ vérifie toujours plus exactement l'équation (11), nous dirons que le schéma aux différences $L_h u^{(h)} = f^{(h)}$ approche le problème aux limites différentiel $Lu = f$ pour la solution u de ce dernier.

S'il y a approximation, on peut estimer que l'équation (14) vérifiée par $[u]_h$ s'obtient de (11) en ajoutant au second membre un terme supplémentaire $\delta f^{(h)}$ peu important (pour h petit). Si la solution $u^{(h)}$ du problème (11) est stable par rapport à la perturbation du second membre $f^{(h)}$, i.e. elle varie peu pour une variation faible du second membre, la différence entre la solution $u^{(h)}$ du problème (11) et la solution $[u]_h$ du problème (14) est donc si faible que l'approximation

$$\delta f^{(h)} \to 0 \quad \text{pour } h \to 0$$

entraîne la convergence

$$u^{(h)} \to [u]_h \quad \text{pour } h \to 0.$$

En vérifiant la convergence (12) nous décomposerons ce problème difficile en deux problèmes plus simples : nous vérifierons d'abord si le problème (1) est vraiment approché par le problème (11) et nous établirons ensuite si le problème (11) est stable. Cette façon d'agir suggère la technique de construction des schémas aux différences convergents pour la résolution numérique du problème (1) : on construit un schéma aux différences approximant ; de tous les procédés d'approximation on choisit celui pour lequel le schéma aux différences est stable.

Ce plan général d'étude de la convergence suppose naturellement les notions rigoureuses d'approximation et de stabilité permettant de démontrer que l'approximation et la stabilité entraînent la convergence. Les définitions de l'approximation et de la stabilité que nous avons esquissées plus haut ne le sont pas. Pour définir l'approximation il faut préciser ce qu'est le résidu $\delta f^{(h)}$ dans le cas général et ce qu'est sa grandeur, et pour définir la stabilité on doit donner un sens précis à la phrase « à une perturbation faible du second membre il correspond une perturbation faible de la solution du problème aux différences $L_h u^{(h)} = f^{(h)}$ ».

La définition rigoureuse de l'approximation et de la stabilité sera le sujet des paragraphes spéciaux.

Exercices

1. Diviser l'intervalle [0, 1] en N intervalles partiels par les points $x_0 = 0, x_1, x_2, \ldots$ $\ldots, x_{N-1}, x_N = 1$, de sorte que

$$\frac{x_{n+1}-x_n}{x_n-x_{n-1}} = q,$$

et éclaircir si on peut utiliser une suite de tels réseaux pour $N \to \infty$ (q étant une constante indépendante de N) en vue de résoudre approximativement le problème

$$\left.\begin{aligned} u'-u &= 0, \\ u(0) &= 1 \end{aligned}\right\}$$

moyennant le schéma aux différences

$$\left.\begin{aligned} \frac{u^{(h)}(x_{n+1})-u^{(h)}(x_n)}{x_{n+1}-x_n} - u^{(h)}(x_n) &= 0 \quad \left(h = \frac{1}{N}\right) \\ u^{(h)}(x_0) &= 1. \end{aligned}\right\}$$

Y a-t-il convergence vers 0 pour $N \to \infty$ du pas $x_{n+1}-x_n$ maximal ?

INDICATION. Discuter de préférence le cas $q > 1$ et s'assurer que

$$\lim_{N\to\infty} u^{(1/N)}(x_N) = \infty.$$

§11. Approximation d'un problème aux limites différentiel par un schéma aux différences

1. Résidu $\delta f^{(h)}$. Attribuons le sens précis à la notion d'approximation du problème aux limites différentiel (1) du §10

$$Lu = f \tag{1}$$

par le schéma aux différences (11) du même paragraphe

$$L_h u^{(h)} = f^{(h)} \tag{2}$$

pour la solution u de (1). A cette fin précisons ce qu'est le résidu $\delta f^{(h)}$

$$L_h[u]_h = f^{(h)} + \delta f^{(h)} \tag{3}$$

apparu par substitution de la fonction discrète $[u]_h$, table de la solution cherchée u, dans l'équation (2), et quelle est sa grandeur.

La tendance de celle-ci vers zéro pour $h \to 0$ sera adoptée par nous comme définition de l'approximation.

Prenons pour commencer un schéma aux différences pour la résolution numérique du problème aux limites différentiel

$$\left.\begin{aligned} \frac{d^2u}{dx^2} + a(x)\frac{du}{dx} + b(x)u &= \cos x, \quad 0 \leqslant x \leqslant 1, \\ u(0) &= 1, \\ u'(0) &= 2. \end{aligned}\right\} \tag{4}$$

Le réseau D_h sera comme par le passé l'ensemble des points $x_n = nh$, $n = 0, 1, \ldots, N$; $h = 1/N$. Acceptons en qualité de schéma aux différences permettant de calculer approximativement $[u]_h$ le système d'égalités

$$\left.\begin{aligned} &\frac{u_{n+1}-2u_n+u_{n-1}}{h^2}+a(x_n)\frac{u_{n+1}-u_{n-1}}{2h}+b(x_n)u_n= \\ &\quad = \cos x_n, \quad n = 1, 2, \ldots, N-1, \\ &u_0 = 1, \\ &\frac{u_1-u_0}{h} = 2, \end{aligned}\right\} \tag{5}$$

résultant de la substitution des dérivées dans (4) par les formules approchées

$$\left.\begin{aligned} \frac{u(x+h)-2u(x)+u(x-h)}{h^2} &\approx \frac{d^2u(x)}{dx^2}, \\ \frac{u(x+h)-u(x-h)}{2h} &\approx \frac{du(x)}{dx}, \\ \frac{u(h)-u(0)}{h} &\approx \frac{du(0)}{dx}. \end{aligned}\right\} \tag{6}$$

En notant

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{n+1}-2u_n+u_{n-1}}{h^2}+a(nh)\dfrac{u_{n+1}-u_{n-1}}{2h}+b(nh)u_n, \\ u_0, \\ \dfrac{u_1-u_0}{h}, \end{cases} \tag{7}$$

$$f^{(h)} \equiv \begin{cases} \cos nh, \\ 1, \\ 2 \end{cases}$$

le schéma aux différences (5) s'écrit sous la forme (2). Le calcul et l'évaluation de la grandeur du résidu $\delta f^{(h)}$ apparu en substituant $[u]_h$ dans l'équation (2) exigent qu'on précise les formules (6).

D'après la formule de Taylor,

$$\begin{aligned} u(x+h) &= u(x)+hu'(x)+\frac{h^2}{2}u''(x)+\frac{h^3}{6}u'''(\xi_1), \\ u(x-h) &= u(x)-hu'(x)+\frac{h^2}{2}u''(x)-\frac{h^3}{6}u'''(\xi_2), \\ u(x+h) &= u(x)+hu'(x)+\frac{h^2}{2}u''(x)+\frac{h^3}{6}u'''(x)+\frac{h^4}{24}u^{(4)}(\xi_3), \\ u(x-h) &= u(x)-hu'(x)+\frac{h^2}{2}u''(x)-\frac{h^3}{6}u'''(x)+\frac{h^4}{24}u^{(4)}(\xi_4), \\ u(x+h) &= u(x)+hu'(x)+\frac{h^2}{2}u''(\xi_5). \end{aligned}$$

Ici ξ_1, ξ_2, ξ_3, ξ_4, ξ_5 sont des points intermédiaires de l'intervalle $[x-h, x+h]$. D'où

$$\left.\begin{aligned}
\frac{u(x+h)-u(x-h)}{2h} &= u'(x)+\frac{h^2}{12}\,[u'''(\xi_1)+u'''(\xi_2)],\\
\frac{u(x+h)-2u(x)+u(x-h)}{h^2} &= u''(x)+\frac{h^2}{24}\,[u^{(4)}(\xi_3)+u^{(4)}(\xi_4)],\\
\frac{u(x+h)-u(x)}{h} &= u'(x)+\frac{h}{2}\,u''(\xi_5).
\end{aligned}\right\} \tag{8}$$

2. Calcul du résidu. Supposons que la solution $u(x)$ du problème (4) admet des dérivées bornées jusqu'à l'ordre 4 inclus. En vertu des formules (8), on peut écrire

$$\frac{u(x+h)-2u(x)+u(x-h)}{h^2}+a(x)\,\frac{u(x+h)-u(x-h)}{2h}+b(x)u(x) =$$
$$= \frac{d^2u(x)}{dx^2}+a(x)\,\frac{du(x)}{dx}+b(x)u(x)+h^2\left[\frac{u^{(4)}(\xi_3)+u^{(4)}(\xi_4)}{24}+\right.$$
$$\left.+a(x)\,\frac{u'''(\xi_1)+u'''(\xi_2)}{12}\right].$$

C'est pourquoi l'expression

$$L_h[u]_h = \begin{cases}
\dfrac{u(x_n+h)-2u(x_n)+u(x_n-h)}{h^2}+a(x_n)\,\dfrac{u(x_n+h)-u(x_n-h)}{2h}+\\
\qquad +b(x_n)\,u(x_n), \quad n = 1, 2, \ldots, N-1,\\
u(0),\\
\dfrac{u(h)-u(0)}{h}
\end{cases}$$

peut se récrire

$$L_h[u]_h = \begin{cases}
\cos x_n+h^2\left[\dfrac{u^{(4)}(\xi_3)+u^{(4)}(\xi_4)}{24}+a(x_n)\times\right.\\
\left.\times\dfrac{u'''(\xi_1)+u'''(\xi_2)}{12}\right], \quad n = 1, 2, \ldots, N-1,\\
1+0,\\
2+h\,\dfrac{u''(\xi_5)}{2}
\end{cases}$$

ou

$$L_h[u]_h = f^{(h)}+\delta f^{(h)},$$

avec

$$\delta f^{(h)} = \begin{cases}
h^2\left[\dfrac{u^{(4)}(\xi_3)+u^{(4)}(\xi_4)}{24}+\dfrac{u'''(\xi_1)+u'''(\xi_2)}{12}\right],\\
0,\\
h\,\dfrac{u''(\xi_5)}{2}.
\end{cases} \tag{9}$$

Il y a intérêt à estimer que $f^{(h)}$ et $\delta f^{(h)}$ définis par (7) et (9) appartiennent à un espace vectoriel normé F_h composé d'éléments de la forme

$$g^{(h)} = \begin{cases} \varphi_n, & n = 1, 2, \ldots, N-1, \\ \psi_0, \\ \psi_1, \end{cases} \tag{10}$$

$\varphi_1, \varphi_2, \ldots \varphi_{n-1}$ ainsi que ψ_0 et ψ_1 étant un système ordonné de nombres ; on peut considérer $g^{(h)}$ comme composé d'une fonction discrète φ_n, $n = 1, 2, \ldots, N-1$, et d'un couple ordonné de nombres ψ_0 et ψ_1. L'addition de deux éléments de l'espace F_h et la multiplication des éléments $g^{(h)}$ par des nombres se font composante par composante. Il est clair que F_h est en l'occurrence un espace vectoriel à $N+1$ dimensions. La norme est prise sur F_h de différentes façons. Si la norme est introduite par l'égalité

$$\|g^{(h)}\|_{F_h} = \max\left(|\psi_0|, |\psi_1|, \max_n |\varphi_n|\right),$$

i.e. si on prend pour norme le maximum des valeurs absolues de toutes les composantes du vecteur $g^{(h)}$, alors on a, conformément à (9),

$$\|\delta f^{(h)}\|_{F_h} \leqslant Ch, \tag{11}$$

C étant une constante fonction de $u(x)$ et indépendante de h.

Il découle de cette inégalité que le résidu $\delta f^{(h)}$ tend vers zéro avec h.

Dans l'équation $L_h u^{(h)} = f^{(h)}$ détaillée par les égalités (5) qui nous a servi d'exemple, L_h peut être assimilé à un opérateur qui fait correspondre à toute fonction discrète $v^{(h)} = \{v_n\}$, $n = 0, 1, \ldots, N$, de l'espace vectoriel de fonctions définies sur le réseau D_h un élément $g^{(h)}$ de la forme (10) de l'espace vectoriel F_h selon la formule

$$L_h v^{(h)} \equiv \begin{cases} \dfrac{v_{n+1} - 2v_n + v_{n-1}}{h^2} + a(x_n)\dfrac{v_{n+1} - v_{n-1}}{2h} + b(x_n)v_n, \\ v_0, \\ \dfrac{v_1 - v_0}{h}. \end{cases}$$

Nous ferons la convention que même dans le cas général du problème aux limites aux différences (2), les seconds membres des équations scalaires écrites dans leur ensemble sous forme d'égalité symbolique

$$L_h u^{(h)} = f^{(h)}$$

sont les composantes du vecteur $f^{(h)}$, élément d'un espace vectoriel normé F_h. On considère alors L_h comme un opérateur qui fait correspondre à toute fonction discrète $u^{(h)} \in U_h$ un élément $f^{(h)} \in F_h$.

Dans ce cas, l'expression $L_h[u]_h$, résultat de l'application de l'opérateur L_h à la fonction discrète $[u]_h$ de U_h et élément de F_h, a un sens.

Le résidu $\delta f^{(h)} = L_h[u]_h - f^{(h)}$ appartient à F_h en tant que différence de deux éléments de cet espace. Sa grandeur est évaluée par $\|\delta f^{(h)}\|_{F_h}$.

3. Approximation de l'ordre de h^k.

DÉFINITION. Nous dirons que le schéma aux différences $L_h u^{(h)} = f^{(h)}$ *approche* le problème $Lu = f$ pour la solution u si $\|\delta f^{(h)}\|_{F_h} \to 0$ lorsque $h \to 0$. Si, de plus, on a l'inégalité

$$\|\delta f^{(h)}\|_{F_h} \leq ch^k,$$

avec $c > 0$ et $k > 0$ des constantes, nous dirons qu'il y a *approximation de l'ordre de h^k* ou d'ordre k en h.

Le fait que u est solution du problème (1) est une information qui peut aider à *construire* le système (2) et à tester l'approximation même. C'est la raison pour laquelle la définition de l'approximation mentionne le problème (1). Il convient néanmoins de souligner que la définition énoncée de l'approximation du problème $Lu = f$ par le schéma aux différences $L_h u^{(h)} = f^{(h)}$ pour la solution u ne s'appuie pas d'elle-même sur l'égalité $Lu = f$ pour la fonction u. On pourrait dire tout simplement que le schéma $L_h u^{(h)} = f^{(h)}$ correspond à l'ordre h^k à u sans éclaircir l'origine de cette dernière fonction. En particulier, si u est solution de deux problèmes différents $L^{(1)}u = f^{(1)}$ et $L^{(2)}u = f^{(2)}$ du type (1) à la fois, le schéma aux différences $L_h u^{(h)} = f^{(h)}$ ou bien approche, ou bien n'approche pas les deux problèmes pour leur solution commune u.

4. Exemples.

EXEMPLE 1. Vu l'estimation (11), l'approximation du problème (4) par le schéma aux différences (5) est du premier ordre en h. Le schéma (5) se prête sans difficulté à un perfectionnement tel que l'approximation soit d'ordre 2. Pour cela notons que toutes les composantes du vecteur $\delta f^{(h)}$ sauf la dernière tendent vers zéro comme h^2 (l'avant-dernière vaut exactement 0).

Seule la dernière composante de $\delta f^{(h)}$, i.e. le résidu obtenu par substitution de $[u]_h$ dans la dernière équation $\frac{u_1 - u_0}{h} = 2$ du système (5), tend vers zéro plus lentement, à savoir comme la première puissance de h. Cette difficulté est facilement aplanie. Selon la formule de Taylor,

$$\frac{u(h) - u(0)}{h} = u'(0) + \frac{h}{2} u''(0) + \frac{h^2}{6} u'''(\xi) =$$

$$= 2 + \frac{h}{2} u''(0) + \frac{h^2}{6} u'''(\xi), \quad 0 < \xi < h.$$

Or, on trouve, à partir de l'équation différentielle (4),

$$u''(0) = -a(0)\,u'(0) - b(0)\,u(0) + \cos 0 = -2a(0) - b(0) + 1.$$

Aussi, en remplaçant la dernière égalité (5) par

$$\frac{u_1 - u_0}{h} = 2 - \frac{h}{2}\,[2a(0) + b(0) - 1], \tag{12}$$

on obtient, au lieu de (7), une autre expression de $f^{(h)}$:

$$f^{(h)} \equiv \begin{cases} \cos x_n, \\ 1, \\ 2-\frac{h}{2}[2a(0)+b(0)-1]. \end{cases}$$

Alors

$$\delta f^{(h)} \equiv \begin{cases} \frac{h^2}{12}\left[\frac{u^{(4)}(\xi_3)+u^{(4)}(\xi_4)}{2}+(u'''(\xi_1)+u'''(\xi_2))\right], \\ 0, \\ \frac{h^2}{6}u'''(\xi) \end{cases}$$

et $\|\delta f^{(h)}\|_{F_h} < C_1 h^2$, où C_1 est une constante indépendante de h. On a donc une approximation d'ordre 2 en h.

Il importe de souligner qu'en construisant la condition aux limites aux différences (12) nous avons utilisé, en plus des conditions aux limites du problème (4), l'équation différentielle même. On peut dire que nous nous sommes servi de la condition aux limites

$$u''(x)+a(x)\,u'(x)+b(x)\,u(x)|_{x=0} = \cos x|_{x=0},$$

découlant de l'équation différentielle.

Exemple 2. Etablissons à quel ordre le schéma aux différences

$$L_h u^{(h)} \equiv \begin{cases} \frac{u_{n+1}-u_{n-1}}{2h}+Au_n = 1+x_n^2, \quad n = 1, 2, \ldots, N-1, \\ u_0 = b, \\ u_1 = b \end{cases} \tag{13}$$

approche le problème

$$\left.\begin{aligned} \frac{du}{dx}+Au &= 1+x^2, \\ u(0) &= b \end{aligned}\right\} \tag{14}$$

pour sa solution u. Un schéma analogue a été analysé au §8 avant que nous ayons introduit une notion rigoureuse d'approximation.

Le rôle de $f^{(h)}$ est ici joué par

$$f^{(h)} = \begin{cases} 1+x_n^2, \quad n = 1, 2, \ldots, N-1, \\ b, \\ b. \end{cases}$$

Ensuite,

$$L_h[u]_h \equiv \begin{cases} \dfrac{u(x_n+h)-u(x_n-h)}{2h}+Au(x_n), & n = 1, \ldots, N-1, \\ u(0), \\ u(h) \end{cases}$$

ou

$$L_h[u]_h \equiv \begin{cases} \left[\dfrac{du(x_n)}{dx}+Au(x_n)\right]+\dfrac{h^2}{6}u'''(\xi_n), & n = 1, \ldots, N-1, \\ u(0), \\ u(0)+h\dfrac{du(\xi_0)}{dx}. \end{cases}$$

Comme la solution $u(x)$ satisfait à l'égalité

$$\frac{du(x_n)}{dx}+Au(x_n) = 1+x_n^2,$$

le résidu $\delta f^{(h)}$ est de la forme

$$\delta f^{(h)} = \begin{cases} \dfrac{h^2}{6}u'''(\xi_n), & n = 1, \ldots, N-1, \\ 0, \\ hu'(\xi_0). \end{cases}$$

L'approximation du problème (14) par le schéma (13) est d'ordre 1 en h. On remarque que les composantes du résidu sont, comme dans l'exemple 1, d'ordres différents en h. L'équation aux différences

$$\frac{u_{n+1}-u_{n-1}}{2h}+Au_n = 1+x_n^2, \quad n = 1, 2, \ldots, N-1, \tag{15}$$

est vérifiée par substitution de $[u]_h$ avec le résidu $\frac{h^2}{6}u'''(\xi_n)$ en h^2, la première condition aux limites

$$u_0 = b \tag{16}$$

l'est exactement et la seconde

$$u_1 = b \tag{17}$$

avec le résidu $hu'(\xi_0)$ de l'ordre de la première puissance de h.

L'erreur d'approximation a été évaluée par

$$\max_{0<x<1} |u'''(x)|, \quad \max_{0<x<1} |u'(x)|.$$

Dans l'exemple considéré, la solution exacte

$$u(x) = u(0)e^{-Ax}+\frac{1+x^2-e^{-Ax}}{A}-\frac{2x}{A^2}$$

permet d'évaluer ces maxima moyennant les données $u(0)$, A du problème :

$$\max_{0<x<1} |u'(x)| = \max_{0<x<1} \left| Au(0)e^{-Ax} + \frac{2x+Ae^{-Ax}}{A} - \frac{2}{A^2} \right| \leqslant$$

$$\leqslant |u(0)||A|(1+e^{-A}) + \frac{2}{|A|} + \frac{2}{A^2} + (1+e^{-A}),$$

$$\max_{0<x<1} |u'''|(x)| = \left| \left[u(0) + \frac{1}{A}\right] A^3 e^{-Ax} \right| \leqslant [|u(0)||A|^3 + |A|^2](1+e^{-A}).$$

Pour des cas plus compliqués on se borne à une estimation grossière de ces dérivées basée sur le théorème de dérivabilité des solutions des équations différentielles ordinaires avec des seconds membres réguliers.

5. Décomposition d'un schéma aux différences en sous-systèmes. Une description détaillée du caractère de l'approximation a exigé que nous parlions non pas de tout le schéma aux différences (13) de la forme (2)

$$L_h u^{(h)} = f^{(h)},$$

mais des sous-systèmes (15), (16), (17). Ces sous-systèmes, dont les deux derniers ne comptent qu'une équation chacun, peuvent s'écrire sous forme d'égalités symboliques respectives

$$l_h^0 u^{(h)} = f_0^{(h)}, \tag{18}$$

$$l_h^1 u^{(h)} = f_1^{(h)}, \tag{19}$$

$$l_h^2 u^{(h)} = f_2^{(h)}. \tag{20}$$

A cet effet on pose

$$l_h^0 u^{(h)} = \frac{u_{n+1} - u_{n-1}}{2h} + Au_n, \quad n = 1, 2, \ldots, N-1,$$

$$l_h^1 u^{(h)} \equiv u_0,$$

$$l_h^2 u^{(h)} \equiv u_1,$$

$$f_0^{(h)} = 1 + x_n^2,$$

$$f_1^{(h)} = b,$$

$$f_2^{(h)} = b.$$

Pour des raisons de commodité et dans le cas général, on décompose souvent le schéma aux différences (2) en deux ou plusieurs sous-systèmes

$$\left.\begin{array}{l} l_h^0 u^{(h)} = f_0^{(h)}, \\ \cdots\cdots\cdots\cdots \\ l_h^r u^{(h)} = f_r^{(h)}, \\ \cdots\cdots\cdots\cdots \\ l_h^R u^{(h)} = f_R^{(h)}, \end{array}\right\} \tag{21}$$

de sorte que

$$L_h u^{(h)} \equiv \begin{cases} l_h^0 u^{(h)}, \\ \cdots\cdots \\ l_h^r u^{(h)}, \\ \cdots\cdots \\ l_h^R u^{(h)}, \end{cases} \qquad f^{(h)} = \begin{cases} f_0^{(h)}, \\ \cdots \\ f_r^{(h)}, \\ \cdots \\ f_R^{(h)}. \end{cases}$$

On a intérêt à prendre le second membre $f_r^{(h)}$ de chaque sous-système $l_h^r u^{(h)} = f_r^{(h)}$ pour élément d'un espace vectoriel normé $F_h^{(r)}$. On munira les espaces F_h et $F_h^{(1)}, F_h^{(2)}, \ldots, F_h^{(R)}$ des normes concordantes telles qu'on ait

$$\|f^{(h)}\|_{F_h} = \max_r \|f_r^{(h)}\|_{F_h^{(r)}}. \tag{22}$$

En partageant (2) en sous-systèmes (21) nous supposerons toujours remplie la condition (22).

Le partitionnement décrit du schéma $L_h u^{(h)} = f^{(h)}$ en (21) permet de parler de l'ordre de correspondance de chaque sous-système pris à part à la solution u du problème (1), $Lu = f$. C'est l'ordre de décroissance de la norme $\|\delta f_r^{(h)}\|_{F_h^{(r)}}$ du résidu $\delta f_r^{(h)}$

$$l_h^{(r)}[u]_h = f_r^{(h)} + \delta f_r^{(h)}$$

pour $h \to 0$. Etant donné le choix concordant des normes (22), tout le schéma aux différences $L_h u^{(h)} = f^{(h)}$ approche le problème $Lu = f$ pour sa solution u à l'ordre égal à celui de décroissance de la norme $\|\delta f^{(r)}\|_{F_h^{(r)}}$ du résidu $\delta f_r^{(h)}$ pour un r qui détermine sa plus lente décroissance.

Si nous décomposons, dans l'exemple 2, le système (13) en sous-systèmes (15) à (17) ou (18) à (20), l'espace $F_h^{(0)}$ est composé des fonctions $f_0^{(h)} = \{f_n\}$ de norme $\|f_0^{(h)}\| = \max_n |f_n|$ définies aux points $x_n = nh$, $n = 1, 2, \ldots$ $\ldots, N-1$, du réseau, et les espaces $F_h^{(1)}$ et $F_h^{(2)}$ unidimensionnels sont formés par des nombres de norme $\|a\| = |a|$. L'équation (18)

$$l_h^{(0)} u^{(h)} = f_0^{(h)}$$

correspond au problème (14) pour la solution u à l'ordre deux, l'équation $l_h^{(1)} u^{(h)} = f_1^{(h)}$ l'est exactement et $l_h^{(2)} u^{(h)} = f_2^{(h)}$ à l'ordre 1. Pour que l'approximation par le schéma (13) soit du deuxième ordre en h, il suffit de « corriger » la condition aux limites $l_h^{(2)} u^{(h)} = b$. Notons que

$$l_h^{(2)}[u]_h = u(h) = u(0) + hu'(0) + \frac{h^2}{2} u''(\xi).$$

Prenons en considération que $u(0) = b$ et que, en vertu de (14),

$$u'(0) = -Au(0) + 1 = -Ab + 1.$$

Posons

$$l_h^{(2)} u^{(h)} = u_1 = b - hAb + h, \quad \text{i.e.} \quad f_2^{(h)} = b - hAb + h,$$

nous aboutirons à la condition

$$l_h^{(2)}[u]_h = u(h) = f_2^{(h)} + O(h^2),$$

i. e. la correspondance de la condition aux limites

$$l_h^{(2)} u^{(h)} = f_2^{(h)} \quad (f_2^{(h)} = b - hAb + h) \tag{23}$$

au problème (14) pour la solution u sera d'ordre 2 en h. Le schéma aux différences (15), (16), (23) approche donc le problème (14) à l'ordre 2 en h.

Le partitionnement de (2) en (21) est conventionnel. On pourrait, par exemple, décomposer le système (13) en deux sous-systèmes dont le premier comprendrait toujours l'équation aux différences (15) et le second les deux conditions aux limites (16) et (17) ou, symboliquement,

$$\left.\begin{aligned} l_h^{(0)} u^{(h)} &= f_0^{(h)}, \\ l_h^{(1)} u^{(h)} &= f_1^{(h)}, \end{aligned}\right\}$$

avec

$$l_h^{(1)} = \begin{cases} u_0, \\ u_1, \end{cases} \quad f_1^{(h)} = \begin{pmatrix} b \\ b \end{pmatrix}.$$

Contrairement au cas du partitionnement (15) à (17) ou (18) à (20), il nous serait impossible d'exprimer brièvement le fait que par substitution de $[u]_h$ la première condition aux limites est remplie exactement et la seconde seulement à l'ordre 1 en h.

6. Remplacement des dérivées par des rapports aux différences. Dans les exemples ci-dessus nous avons obtenu des schémas aux différences en remplaçant les dérivées de l'équation différentielle par des rapports aux différences. Ce procédé fort répandu permet de construire, pour tout problème aux limites différentiel possédant une solution $u(x)$ suffisamment régulière, un schéma aux différences d'ordre d'approximation donné d'avance.

Montrons, en effet, que la dérivée $d^k u/dx^k$, avec k quelconque, peut être remplacée par un rapport aux différences de façon que, $u(x)$ étant suffisamment régulière, l'erreur commise soit d'ordre p quelconque donné d'avance par rapport au pas h du réseau aux différences. Utilisons la technique des coefficients indéterminés.

Ecrivons une égalité de la forme

$$\frac{d^k(x)}{dx^k} = h^{-k} \sum_{s=-s_1}^{s_2} a_s u(x+sh) + O(h^p) \tag{24}$$

et choisissons des coefficients indéterminés a_s, $s = -s_1, -s_1+1, \ldots, s_2$, indépendants de h qui vérifient (24). Les limites de sommation $s_1 \geqslant 0$ et $s_2 \geqslant 0$ peuvent être quelconques mais telles que l'ordre s_1+s_2 du rapport aux différences $h^{-k} \sum a_s u(x+sh)$ satisfasse à l'inégalité $s_1+s_2 \geqslant k+p-1$. Selon la formule de Taylor,

$$u(x+sh) = u(x) + sh\frac{du(x)}{dx} + \frac{(sh)^2}{2!}\frac{d^2u(x)}{dx^2} + \ldots$$

$$\ldots + \frac{(sh)^{k+p-1}}{(k+p-1)!}\frac{d^{k+p-1}u(x)}{dx^{k+p-1}} + \frac{(sh)^{k+p}}{(k+p)!}\frac{d^{k+p}u(\xi)}{dx^{k+p}}.$$

Substituons cette expression à $u(x+sh)$ de (24) et réduisons les termes semblables, il vient

$$\frac{d^k u(x)}{dx^k} = h^{-k}\Big[u(x)\sum a_s + \frac{du(x)}{dx}\,\frac{h}{1!}\sum s a_s + \ldots + \frac{d^{k+p-1}u(x)}{dx^{k+p-1}}\times$$
$$\times\frac{h^{k+p-1}}{(k+p-1)!}\sum s^{k+p-1}a_s\Big] + \frac{h^p}{(k+p)!}\sum s^{k+p}a_s\,\frac{d^{k+p}u(\xi_s)}{dx^{k+p}}\,.$$

En égalant les coefficients des mêmes puissances de h^s, $s = -k, -k+1, \ldots, p-1$, aux deux membres de cette égalité, nous obtenons le système d'équations en a_s :

$$\left.\begin{array}{l}\sum a_s = 0,\\ \sum s a_s = 0,\\ \ldots\ldots\ldots\ldots\\ \sum s^{k-1}a_s = 0,\\ \sum s^k a_s = k!,\\ \sum s^{k+1}a_s = 0,\\ \ldots\ldots\ldots\ldots\\ \sum s^{k+p-1}a_s = 0.\end{array}\right\} \qquad (25)$$

Si $s_1+s_2 = k+p-1$, les $k+p$ égalités écrites constituent un système linéaire en autant d'inconnues a_s. Son déterminant (déterminant de Vandermonde)

$$\begin{vmatrix} 1 & 1 & \ldots & 1\\ -s_1 & -s_1+1 & \ldots & s_2\\ \ldots & \ldots & \ldots & \ldots\\ (-s_1)^{k+p-1} & (-s_1+1)^{k+p-1} & \ldots & s_2^{k+p-1}\end{vmatrix}$$

est différent de zéro. Il existe donc un seul jeu des coefficients a_s vérifiant le système (25). Si $s_1+s_2 \geqslant k+p$, il en existe évidemment plus d'un.

Ainsi, il existe un seul rapport aux différences du premier ordre de la forme

$$h^{-1}[a_0u(x)+a_1u(x+h)]$$

approchant du/dx à l'ordre 1 en h. Il s'obtient pour

$$\frac{du}{dx} = \frac{u(x+h)-u(x)}{h} + O(h).$$

De même, il existe un seul rapport aux différences d'ordre 1 de la forme

$$h^{-1}[a_{-1}u(x-h)+a_0u(x)],$$

qui approche du/dx à l'ordre 1 en h :

$$\frac{du}{dx} = \frac{u(x)-u(x-h)}{h} + O(h).$$

Parmi des rapports aux différences du second ordre de la forme

$$h^{-1}[a_{-1}u(x-h)+a_0u(x)+a_1u(x+h)]$$

il y en a une infinité approchant du/dx à l'ordre 1 en h et un seul qui le fait à l'ordre 2. En résolvant le système (25) pour ce cas on voit que, lorsque $a_1 = {}^1/_2$, $a_0 = 0$, $a_{-1} = -{}^1/_2$,

$$\frac{du}{dx} = \frac{u(x+h)-u(x-h)}{2h} + O(h^2).$$

En voulant obtenir une approximation de d^2u/dx^2 qui soit d'ordre 2, on a $k = 2$, $p = 2$, et il faut que $s_1+s_2 \geqslant 3$. Aussi ne peut-on utiliser en l'occurrence qu'un seul des rapports aux différences

$$h^{-2}(a_{-1}u(x-h)+a_0u(x)+a_1u(x+h)+a_2u(x+2h)). \tag{26}$$

La résolution du système (26) donne

$$a_{-1} = a_1 = 1, \quad a_0 = -2, \quad a_2 = 0,$$

i.e. l'égalité que nous avons utilisée à plusieurs reprises

$$\frac{d^2u(x)}{dx^2} = \frac{u(x+h)-2u(x)+u(x-h)}{h^2}+O(h^2).$$

7. D'autres méthodes de construction de schémas aux différences. La technique de construction de schémas aux différences décrite au numéro précédent n'est pas l'unique ni souvent la meilleure. Le paragraphe 19 traitera de certains autres procédés conduisant à des schémas aux différences parmi les plus usités. Pour le moment nous allons nous borner à un exemple.

Le schéma aux différences simple

$$L_hu^{(h)} \equiv \begin{cases} \dfrac{u_{n+1}-u_n}{h}-G(x_n, u_n) = 0, & n = 0, 1, \ldots, N-1, \\ u_0 = a \end{cases}$$

dit *schéma d'Euler* approche le problème

$$\left.\begin{aligned} \frac{du}{dx}-G(x, u) = 0, \quad 0 \leqslant x \leqslant 1, \\ u(0) = a \end{aligned}\right\} \tag{27}$$

à l'ordre 1 en h. Connaissant u_n, la valeur u_{n+1} se calcule d'après la formule $u_{n+1} = u_n+hG(x_n, u_n)$. Le schéma

$$L_hu^{(h)} \equiv \begin{cases} \dfrac{u_{n+1}-u_n}{h}-\dfrac{1}{2}[G(x_n, u_n)+G(x_{n+1}, \bar{u})] = 0, \\ u_0 = a, \end{cases}$$

où $\bar{u} = u_n+hG(x_n, u_n)$, s'appelle *schéma prédicteur-correcteur d'Euler*. C'est un des schémas de Runge-Kutta d'ordre d'approximation deux dont nous parlerons en détail au §19. Si u_n est déjà calculé, on obtient, d'après le schéma d'Euler,

$$\bar{u} = u_n+hG(x_n, u_n)$$

puis on trouve la valeur corrigée de $\bar{u}$

$$u_{n+1} = u_n+\frac{h}{2}[G(x_n, u_n)+G(x_{n+1}, \bar{u})].$$

Exercice

Vérifier que le schéma prédicteur-correcteur d'Euler approche le problème (27) pour la solution régulière $u(x)$ à l'ordre deux en h.

§12. Définition de la stabilité d'un schéma aux différences. Convergence en tant que conséquence de l'approximation et de la stabilité

1. Définition de la stabilité. Etant donné le problème aux limites différentiel

$$Lu = f, \tag{1}$$

on calcule approximativement sa solution u en construisant un schéma aux différences

$$L_h u^{(h)} = f^{(h)} \tag{2}$$

qui approche ce problème pour u à un certain ordre k en h. Cela signifie que le résidu $\delta f^{(h)}$

$$L_h[u]_h = f^{(h)} + \delta f^{(h)},$$

apparu en substituant la table $[u]_h$ de la solution u dans (2), vérifie une estimation de la forme

$$\|\delta f^{(h)}\|_{F_h} \leqslant C_1 h^k, \tag{3}$$

avec C_1 une constante indépendante de h. On vérifie aisément que le schéma aux différences

$$L_h u^{(h)} \equiv \begin{cases} 4\,\dfrac{u_{n+1}-u_{n-1}}{2h} - 3\,\dfrac{u_{n+1}-u_n}{h} + Au_n = 0, & n = 1, 2, \ldots, N-1, \\ u_0 = b \end{cases}$$

approche

$$\frac{du}{dx} + Au = 0,$$
$$u(0) = b$$

pour u à l'ordre 1 en h. Comme montré §9, la solution $u^{(h)}$ fournie par ce schéma ne tend cependant pas vers $[u]_h$ pour $h \to 0$.

Ainsi, l'approximation ne suppose en général pas la convergence. Pour qu'il y ait convergence, il faut encore la stabilité.

DÉFINITION 1. Le schéma aux différences (2) sera dit *stable* s'il existe des nombres $h_0 > 0$ et $\delta > 0$ tels que, quels que soient $h < h_0$ et $\varepsilon^{(h)} \in F_h$, $\|\varepsilon^{(h)}\|_{F_h} < \delta$, le problème aux différences

$$L_h z^{(h)} = f^{(h)} + \varepsilon^{(h)}, \tag{4}$$

obtenu de (2) par adjonction de la perturbation $\varepsilon^{(h)}$ au second membre, admette une, et une seule, solution $z^{(h)}$, la différence entre celle-ci et la solution $u^{(h)}$ du problème non perturbé (2) étant une fonction $z^{(h)} - u^{(h)}$ définie sur un réseau et évaluée par

$$\|z^{(h)} - u^{(h)}\|_{U_h} \leqslant C\,\|\varepsilon^{(h)}\|_{F_h}, \tag{5}$$

où C est une constante indépendante de h.

En particulier, l'inégalité (5) signifie qu'une petite perturbation $\varepsilon^{(h)}$ du second membre du schéma aux différences (2) induit uniformément en h une petite perturbation $z^{(h)}-u^{(h)}$ sur la solution.

Supposons que l'opérateur L_h de U_h dans F_h est linéaire. La définition ci-dessus de la stabilité est alors équivalente à la

DÉFINITION 2. Le schéma aux différences (2) d'opérateur linéaire L_h sera dit *stable* si, quel que soit $f^{(h)} \in F_h$, l'équation $L_h u^{(h)} = f^{(h)}$ admet une seule solution $u^{(h)} \in U_h$ et si de plus

$$\|u^{(h)}\|_{U_h} \leqslant C\|f^{(h)}\|_{F_h}, \tag{6}$$

avec C une constante indépendante de h.

Démontrons l'équivalence des deux définitions pour l'opérateur linéaire L_h.

Commençons par établir que la stabilité au sens de la définition 2 du schéma aux différences (2) entraîne la stabilité au sens de la définition 1. Admettons que le problème linéaire (2) possède, pour tous les $h < h_0$ considérés et pour $f^{(h)} \in F_h$ arbitraire, une solution unique et qu'on a l'estimation (6). En retranchant (2) de (4) nous obtenons

$$L_h(z^{(h)}-u^{(h)}) = \varepsilon^{(h)},$$

d'où, en vertu de (6), l'estimation (5) pour $\varepsilon^{(h)} \in F_h$ arbitraire et donc la stabilité au sens de la définition 1.

Montrons maintenant que la stabilité au sens de la définition 1 entraîne celle au sens de la définition 2. Vu la définition 1, pour certains $h_0 > 0$, $\delta > 0$ et pour $h < h_0$ et $\varepsilon^{(h)} \in F_h$, $\|\varepsilon^{(h)}\|_{F_h} < \delta$, quelconques, les équations

$$L_h z^{(h)} = f^{(h)} + \varepsilon^{(h)},$$
$$L_h u^{(h)} = f^{(h)}$$

admettent chacune une solution et une seule. Posons $w^{(h)} \equiv z^{(h)}-u^{(h)}$ et retranchons ces égalités terme à terme, il vient

$$L_h w^{(h)} = \varepsilon^{(h)},$$

de plus, en vertu de (5),

$$\|w^{(h)}\|_{U_h} \leqslant C\|\varepsilon^{(h)}\|_{F_h}.$$

Il est évident qu'à condition de modifier les notations de la solution et du second membre de $L_n w^{(h)} = \varepsilon^{(h)}$, le dernier résultat peut s'énoncer comme suit : pour $h < h_0$ et $f^{(h)} \in F_h$, $\|f^{(h)}\|_{F_h} < \delta$, quelconques, le problème (2) admet une solution unique $u^{(h)}$. Celle-ci vérifie l'estimation (6). Or, s'il en est ainsi, notre affirmation sur l'existence et l'unicité de $u^{(h)}$ et sur l'estimation (6) est vraie non seulement pour tous les $f^{(h)}$ tels que $\|f^{(h)}\|_{F_h} < \delta$, mais en général pour tous les $f^{(h)} \in F_h$, i.e. il y a stabilité au sens de la définition 2.

En effet, soit $\|f^{(h)}\|_{F_h} \geqslant \delta$. Démontrons l'existence d'une solution,

son unicité et l'estimation (6). Posons

$$u^{(h)} = \frac{2\,\|f^{(h)}\|_{F_h}}{\delta}\,\tilde{u}^{(h)}, \quad f^{(h)} = \frac{2\,\|f^{(h)}\|_{F_h}}{\delta}\,\tilde{f}^{(h)}.$$

Nous obtenons pour $\tilde{u}^{(h)}$

$$L_h\tilde{u}^{(h)} = \tilde{f}^{(h)},$$

de plus

$$\|\tilde{f}^{(h)}\|_{F_h} = \frac{\delta}{2\,\|f^{(h)}\|_{F_h}}\,\|f^{(h)}\|_{F_h} = \frac{\delta}{2} < \delta.$$

Aussi l'équation $L_h\tilde{u}^{(h)} = \tilde{f}^{(h)}$ n'admet-elle qu'une solution unique, en outre

$$\|\tilde{u}^{(h)}\|_{U_h} \leqslant C\,\|\tilde{f}^{(h)}\|_{F_h}.$$

En vertu des formules donnant la relation entre $u^{(h)}$ et $\tilde{u}^{(h)}$, ainsi qu'entre $f^{(h)}$ et $\tilde{f}^{(h)}$, il en résulte l'existence d'une solution de (2) et son unicité ainsi que l'estimation (6) pour $f^{(h)} \in F_h$ considéré arbitraire.

2. Relation entre l'approximation, la stabilité et la convergence. Démontrons maintenant que l'approximation et la stabilité impliquent la convergence.

THÉORÈME. *Si le schéma aux différences $L_hu^{(h)} = f^{(h)}$ approche le problème $Lu = f$ pour la solution u à l'ordre k en h et est stable, alors la solution $u^{(h)}$ du problème aux différences $L_hu^{(h)} = f^{(h)}$ converge vers $[u]_h$ et on a l'estimation*

$$\|[u]_h - u^{(h)}\|_{U_h} \leqslant (CC_1)h^k, \tag{7}$$

C et C_1 étant les nombres figurant dans les estimations (3) *et* (5).

DÉMONSTRATION. Posons $\varepsilon^{(h)} \equiv \delta f^{(h)}$, $[u]_h \equiv z^{(h)}$. L'estimation (5) prend alors la forme

$$\|[u]_h - u^{(h)}\|_{U_h} \leqslant C\,\|\delta f^{(h)}\|_{F_h}.$$

Compte tenu de (3), il en résulte de suite l'inégalité à démontrer (7).

A titre d'illustration démontrons la stabilité du schéma aux différences d'Euler

$$\left.\begin{aligned} \frac{u_{n+1}-u_n}{h} - G(x_n, u_n) &= \varphi_n, \quad n = 0, 1, \ldots, N-1, \\ u_0 &= \psi, \end{aligned}\right\} \tag{8}$$

$x_n = nh$, $h = 1/N$, utilisé pour la résolution numérique du problème aux limites différentiel

$$\left.\begin{aligned} \frac{du}{dx} - G(x, u) &= \varphi(x), \quad 0 \leqslant x \leqslant 1, \\ u_0 &= \psi. \end{aligned}\right\} \tag{9}$$

Soient $G(x, u)$ et $\varphi(x)$ tels qu'il existe une solution $u(x)$ à dérivée seconde bornée. Nous admettrons de plus que $G(x, u)$ a sa dérivée par rapport à u

bornée

$$\left|\frac{\partial G}{\partial u}\right| < M. \tag{10}$$

Le lecteur est prié de vérifier que le schéma aux différences (8) approche le problème (3) pour la solution $u(x)$ à l'ordre 1 en h. (L'équation aux différences correspond au problème au premier ordre et la condition aux limites $u_0 = \psi$ exactement.) Définissons les normes

$$\|u^{(h)}\|_{U_h} = \max_n |u_n|, \quad \|f^{(h)}\|_{F_h} = \max\left\{|\psi|, \max_m |\varphi(x_m)|\right\}$$

et testons la stabilité du schéma (8) que nous mettons sous la forme (2) en posant

$$L_h u^{(h)} = \begin{cases} \frac{u_{n+1}-u_n}{h} - G(x_n, u_n), & n = 0, 1, \ldots, N-1, \\ u_0, \end{cases}$$

$$f^{(h)} = \begin{cases} \varphi(x_n), & n = 0, 1, \ldots, N-1, \\ \psi. \end{cases}$$

Le problème

$$L_h z^{(h)} = f^{(h)} + \varepsilon^{(h)}$$

s'écrit sous forme détaillée

$$\left.\begin{aligned} \frac{z_{n+1}-z_n}{h} - G(x_n, z_n) &= \varphi(x_n) + \varepsilon_n, \quad n = 0, 1, \ldots, N-1, \\ z_0 &= \psi + \varepsilon, \end{aligned}\right\} \tag{11}$$

où

$$\varepsilon^{(h)} = \begin{cases} \varepsilon_n, & n = 0, 1, \ldots, N-1, \\ \varepsilon. \end{cases}$$

Retranchons terme à terme des équations (11) les équations (8) correspondantes. Notons

$$z_n - u_n = w_n$$

et prenons en considération l'égalité

$$G(x_n, z_n) - G(x_n, u_n) = \frac{\partial G(x_n, \xi_n)}{\partial u} w_n = M_n^{(h)} w_n,$$

où ξ_n est un nombre compris entre z_n et u_n. Nous obtenons le système d'équations en $w^{(h)} = (w_0, w_1, \ldots, w_n, \ldots, w_N)$:

$$\left.\begin{aligned} \frac{w_{n+1}-w_n}{h} - M_n^{(h)} w_n &= \varepsilon_n, \quad n = 0, 1, \ldots, N-1, \\ w_0 &= \varepsilon. \end{aligned}\right\} \tag{12}$$

Etant donné que $M_n^{(h)} \leqslant M$ en vertu de (10) et que $nh \leqslant Nh = 1$, nous

avons

$$\begin{aligned}
|w_{n+1}| &= |(1+hM_n^{(h)})w_n+h\varepsilon_n| \leqslant (1+Mh)|w_n|+h|\varepsilon_n| \leqslant \\
&\leqslant (1+Mh)^2|w_{n-1}|+h(1+Mh)|\varepsilon_{n-1}|+h|\varepsilon_n| \leqslant \\
&\leqslant (1+Mh)^2|w_{n-1}|+2h(1+Mh)\|\varepsilon^{(h)}\|_{F_h} \leqslant \\
&\leqslant (1+Mh)^3|w_{n-2}|+3h(1+Mh)^2\|\varepsilon^{(h)}\|_{F_h} \leqslant \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\leqslant (1+Mh)^{n+1}|w_0|+(n+1)h(1+Mh)^n\|\varepsilon^{(h)}\|_{F_h} \leqslant \\
&\leqslant (1+Mh)^{n+1}\|\varepsilon^{(h)}\|_{F_h}+(1+Mh)^n\|\varepsilon^{(h)}\|\,F_h \leqslant \\
&\leqslant 2(1+Mh)^N\|\varepsilon^{(h)}\|_{F_h} \leqslant 2e^M\|\varepsilon^{(h)}\|_{F_h}.
\end{aligned}$$

L'inégalité démontrée

$$|w_{n+1}| \leqslant 2e^M\|\varepsilon^{(h)}\|_{F_h}$$

entraîne une estimation de la forme (6)

$$\|w^{(h)}\|_{U_h} \leqslant 2e^M\|\varepsilon^{(h)}\|_{F_h},$$

qui signifie la stabilité avec une constante $C = 2e^M$. En vertu du théorème, le schéma aux différences (8) est convergent à l'ordre 1 en h.

Etudions la convergence du schéma aux différences (7) §10

$$\left.\begin{aligned}
&\frac{u_{n+1}-2u_n+u_{n-1}}{h^2}-(1+x_n^2)u_n = \sqrt{1+x_n}, \quad n = 1, 2, \ldots, N-1, \\
&u_0 = 2, \\
&u_N = 1
\end{aligned}\right\} \quad (13)$$

pour le problème aux limites différentiel (4) du même paragraphe. Grâce à la formule

$$\frac{u(x+h)-2u(x)+u(x-h)}{h^2} = u''(x)+\frac{h^2}{12}u^{(4)}(\xi),$$

il est évident que le problème (13) approche le problème (4) §10 à l'ordre deux en h(*). Vérifions s'il y a stabilité. Le problème considéré étant linéaire, on teste la stabilité en établissant l'existence de la solution du problème

$$\left.\begin{aligned}
&\frac{u_{n+1}-2u_n+u_{n-1}}{h^2}-(1+x_n^2)u_n = g_n, \quad n = 1, 2, \ldots, N-1, \\
&u_0 = \alpha, \\
&u_N = \beta,
\end{aligned}\right\} \quad (14)$$

(*) Notons que s'il s'agissait de l'équation

$$u''-(1+x^2)u = \sqrt{x},$$

différant peu de l'équation étudiée, on ne pourrait conclure à l'approximation puisque $|u'''(x)|$ ne serait pas borné (démontrer rigoureusement le caractère non borné de $u'''(x)$).

quels que soient $\{g_n\}$, α et β, et en obtenant l'estimation

$$\max_n |u_n| \leqslant C\left(\max_n |g_n|,\ |\alpha|,\ |\beta|\right). \tag{15}$$

Un problème de la forme (14) a été étudié au §4 (p. 40). Sous l'hypothèse

$$|b_n| \geqslant |a_n|+|c_n|+\delta, \quad \delta > 0,$$

nous y avons démontré l'existence de la solution et l'estimation

$$|u_n| < \max\left[|\alpha|,\ |\beta|,\ \frac{1}{\delta}\max_m |g_m|\right] \tag{16}$$

pour un problème de la forme

$$\left.\begin{aligned} a_n u_{n-1}+b_n u_n+c_n u_{n+1} &= g_n, \\ u_0 &= \alpha, \\ u_N &= \beta. \end{aligned}\right\}$$

S'agissant du problème (14) on a

$$a_n = \frac{1}{h^2},\ c_n = \frac{1}{h^2},\ |b_n| = \frac{2}{h^2}+1+x_n^2 > |a_n|+|c_n|+1.$$

C'est pourquoi l'estimation (16) entraîne l'estimation (15) avec une constante $C = 1$. La stabilité est démontrée.

La circonstance suivante pourrait s'avérer utile pour démontrer la convergence en vérifiant l'approximation et la stabilité.

Supposons le schéma (2) décomposé en

$$l_h^{(0)} u^{(h)} = f_0^{(h)}, \tag{17}$$

$$l_h^{(1)} u^{(h)} = f_1^{(h)}, \tag{17'}$$

de sorte que

$$L_h u^{(h)} \equiv \begin{cases} l_h^{(0)} u^{(h)}, \\ l_h^{(1)} u^{(h)}, \end{cases}$$

$$f^{(h)} = \begin{cases} f_0^{(h)}, \\ f_1^{(h)}, \end{cases} \quad \delta f^{(h)} = \begin{cases} \delta f_0^{(h)}, \\ \delta f_1^{(h)}. \end{cases}$$

Supposons ensuite que l'approximation du problème (1) par le schéma aux différences (2) est d'ordre k en h, i.e. on a la condition (3), et que le sous-système (17′) correspond exactement à (1) pour la solution u, i.e. $\delta f_h^{(1)} = 0 \in F_h^{(1)}$:

$$\delta f^{(h)} = \begin{cases} \delta f_0^{(h)}, \\ 0. \end{cases} \tag{18}$$

Pour que la solution $u^{(h)}$ du problème (2) converge alors vers la fonction cherchée $[u]_h$, i.e. pour la validité de l'estimation (7), il suffit que l'estimation (5) ait lieu non pas pour $\varepsilon^{(h)} \in F_h$ quelconques, mais seulement pour

tous les $\varepsilon^{(h)}$ de la forme

$$\varepsilon^{(h)} = \begin{cases} \varepsilon_0^{(h)}, \\ 0, \end{cases} \tag{19}$$

où $0 \in F_h^{(1)}$.

La démonstration calque celle du théorème de la convergence (voir plus haut).

Le lecteur vérifiera aisément que dans le cas de l'opérateur linéaire L_h l'exigence que l'estimation (5) n'ait lieu que pour tous les $\varepsilon^{(h)}$ de la forme (19) est satisfaite en même temps que la condition imposant à l'estimation (6) d'être vraie pour tous les $f^{(h)}$ de la même forme spéciale

$$f^{(h)} = \begin{cases} f_0^{(h)}, \\ 0, \end{cases}$$

où $0 \in F_h^{(1)}$.

La démonstration de la convergence du schéma aux différences (13) aurait pu, par exemple, utiliser le fait que les deux conditions aux limites

$$l_h^{(1)} u^{(h)} \equiv \begin{Bmatrix} u_0 = 2, \\ u_N = 1 \end{Bmatrix} \equiv f_h^{(1)}$$

sont remplies exactement :

$$l_h^{(1)}[u]_h = \begin{cases} u(0) = 2, \\ u(1) = 1, \end{cases}$$

en y portant la table de la solution $[u]_h$ du problème (4) §10.

On pourrait donc vérifier l'inégalité (15) ou, autrement dit, la stabilité du schéma aux différences (13), non pas pour un second membre quelconque

$$f^{(h)} = \begin{cases} g_n, & n = 1, 2, \ldots, N-1, \\ \alpha, \\ \beta, \end{cases}$$

mais pour des seconds membres de la forme

$$f^{(h)} = \begin{cases} g_n, & n = 1, 2, \ldots, N-1, \\ 0, \\ 0 \end{cases}$$

lorsque $\alpha = 0$ et $\beta = 0$.

En ce qui concerne le problème (13), nous n'avons pas mis à profit cette circonstance simplificatrice. Elle s'avérera par contre utile dans certains problèmes plus compliqués (pour des équations aux dérivées partielles).

Disons pour conclure que le schéma de démonstration de la convergence de la solution de $L_h u^{(h)} = f^{(h)}$ vers celle de $Lu = f$ par vérification de l'approximation et de la stabilité a un caractère général. On entend par $Lu = f$ toute équation fonctionnelle et pas obligatoirement un problème aux limites

pour une équation différentielle ordinaire. La nature du problème dont u est solution est sans importance. L'équation $Lu = f$ ne sert qu'à construire l'équation aux différences $L_h u^{(h)} = f^{(h)}$. Développons cette idée au numéro suivant.

3. Schéma aux différences convergent pour une équation intégrale. Construisons et étudions un schéma aux différences pour calculer la solution de l'équation intégrale

$$Lu \equiv u(x) - \int_0^1 K(x, y)\, u(y)\, dy = f(x).$$

Nous supposerons que $|K(x, y)| < \varrho < 1$.

Définissons N, posons $h = 1/N$ et cherchons la table $[u]_h$ des valeurs de la solution sur le réseau $x_n = nh$, $n = 0, 1, \ldots, N$. Le schéma aux différences s'obtient en remplaçant approximativement selon la formule de quadrature des trapèzes l'intégrale de

$$u(x_n) - \int_0^1 K(x_n, y)\, u(y)\, dy = f(x_n), \qquad n = 0, 1, \ldots, N,$$

par une somme. Rappelons cette formule : une fonction $\varphi(y)$ deux fois dérivable sur le segment $0 \leqslant y \leqslant 1$ vérifie l'égalité approchée

$$\int_0^1 \varphi(y)\, dy \approx h\left(\frac{\varphi_0}{2} + \varphi_1 + \varphi_2 + \ldots + \varphi_{N-1} + \frac{\varphi_N}{2}\right), \qquad h = \frac{1}{N},$$

l'erreur étant $C(h^2)$. Par substitution mentionnée, nous trouvons

$$u_n^{(h)} - h\left(\frac{K(x_n, 0)}{2}\, u_0^{(h)} + K(x_n, h) u_1^{(h)} + \ldots + K(x_n, y_{N-1}) u_{N-1}^{(h)} + \right.$$
$$\left. + \frac{K(x_n, 1)}{2}\, u_N^{(h)}\right) = f(x_n), \qquad n = 0, 1, \ldots, N. \tag{20}$$

Ce système d'égalités se met sous la forme $L_h u^{(h)} = f^{(h)}$ à condition de poser

$$L_h u^{(h)} = \begin{cases} g_0, \\ g_1, \\ \cdots \\ g_N, \end{cases} \qquad f^{(h)} = \begin{cases} f(0), \\ f(h), \\ \cdots \\ f(1), \end{cases}$$

où

$$g_n = u_n^{(h)} - h\left[\frac{K(x_n, 0)}{2}\, u_0^{(h)} + K(x_n, h) u_1^{(h)} + \ldots + \frac{K(x_n, 1)}{2}\, u_N^{(h)}\right], \qquad n = 0, 1, \ldots, N.$$

Le schéma aux différences construit $L_h u^{(h)} = f^{(h)}$ approche le problème $Lu = f$ pour la solution u à l'ordre deux en h puisque la formule de quadrature des trapèzes est exacte à l'ordre 2. Vérifions la stabilité. Soient $u^{(h)} = (u_0, u_1, \ldots, u_N)$ une solution du système (20) et u_s une des composantes de la solution qui sont au moins égales aux autres en module :

$$|u_s| = \max_m |u_m|.$$

Il résulte de l'équation de numéro $n = s$ du système (20) :

$$|f(x_s)| = \left| u_s^{(h)} - h\left(\frac{K(x_s, 0)}{2}\, u_0^{(h)} + K(x_s, h) u_1^{(h)} + \ldots + \frac{K(x_s, 1)}{2}\, u_N^{(h)}\right)\right| \geqslant$$
$$\geqslant |u_1^{(h)}| - h\left(\frac{\varrho}{2} + \varrho + \ldots + \varrho + \frac{\varrho}{2}\right) |u_s^{(h)}| = (1 - Nh\varrho)\, |u_s^{(h)}| = (1 - \varrho)\, |u_s^{(h)}|.$$

Donc

$$\|u^{(h)}\|_{U_h} = \max_m |u_m^{(h)}| = |u_s^{(h)}| \leqslant \frac{1}{1-\varrho}|f(x_s)| \leqslant \frac{1}{1-\varrho}\|f^{(h)}\|_{F_h}. \tag{21}$$

Il en découle en particulier pour $f(x_m) \equiv 0$ que le système (20) n'admet pas de solutions non triviales et qu'il est donc résoluble de façon unique quel que soit le second membre $f^{(h)}$. L'inégalité (21) signifie la stabilité (6) avec une constante $C = 1/(1-\varrho)$. En vertu du théorème de la convergence, la solution $u^{(h)}$ du problème $L_h u^{(h)} = f^{(h)}$ satisfait à l'inégalité

$$\|[u]_h - u^{(h)}\|_{U_h} = \max_m |u(mh) - u_m^{(h)}| \leqslant Ah^2,$$

A étant une constante.

§13. Choix des normes

Les notions de convergence, d'approximation et de stabilité introduites aux §§10-12 ont un sens si l'on met d'une façon ou d'une autre des normes sur les espaces U_h et F_h contenant respectivement la solution $u^{(h)}$ et le second membre $f^{(h)}$ du schéma aux différences

$$L_h u^{(h)} = f^{(h)}$$

utilisé pour le calcul approché de la solution u du problème aux limites différentiel $Lu = f$.

Voyons à quel point le choix des normes dans les espaces U_h et F_h peut être arbitraire et considérons en premier lieu la norme $\|\cdot\|_{U_h}$ dont la grandeur évalue l'écart entre la solution approchée $u^{(h)}$ et la fonction $[u]_h$ définie sur le réseau, i.e. la table de valeurs de la solution u. Dans tous les exemples précédents nous avons utilisé une norme définie par l'égalité

$$\|z^{(h)}\|_{U_h} = \max_k |z_k^{(h)}|. \tag{1}$$

Le maximum est pris en tous les points du réseau D_h, domaine de définition de la fonction $z^{(h)} \in U_h$. On pourrait certes poser

$$\|z^{(h)}\|_{U_h} = h \max_k |z_k^{(h)}|, \tag{2}$$

ou

$$\|z^{(h)}\|_{U_h} = \frac{1}{h} \max_k |z_k^{(h)}|, \tag{3}$$

ou même

$$\|z^{(h)}\|_{U_h} = 2^{-1/h} \max_k |z_k^{(h)}|.$$

La dernière norme paraît commode puisqu'on constate pour elle la convergence du schéma aux différences

$$\left.\begin{aligned} 4\frac{u_{n+1}-u_{n-1}}{2h} - 3\frac{u_{n+1}-u_n}{h} + Au_n &= 0, \quad n = 0, 1, \ldots, N-1, \\ u_0 &= a, \\ u_1 &= ae^{-Ah} \end{aligned}\right\}$$

pour la résolution du problème

$$\left.\begin{aligned} u'+Au &= 0, \quad 0 \leqslant x \leqslant 1, \\ u(0) &= a, \end{aligned}\right\}$$

schéma construit au §9 en tant qu'exemple d'un schéma à rejeter. En effet, en vertu de l'égalité

$$u(nh)-u_n = ae^{-Ax_n}-u_n = -\left[\frac{3}{2}A^2ae^{Ax_n}+O(h)\right]h^2 2^{x_n/h}+O(h)$$

découlant de la relation (7) §9, la quantité

$$\|[u]_h-u^{(h)}\|_{U_h} = 2^{-1/h}\max_m |u(mh-u_m^{(h)})|$$

tend vers zéro avec le passage à un réseau plus fin. Il est cependant clair que cette tendance ne signifie en aucun sens raisonnable qu'il en est de même de l'erreur $z^{(h)} = [u]_h-u^{(h)}$ vu que les différences $u(nh)-u_n$ peuvent croître brusquement (presque comme $2^{1/h}$), ce qui est justement notre cas.

Les normes (2) et (3) ne sont pas non plus à recommander parce qu'elles ne caractérisent pas suffisamment l'erreur $[u]_h-u^{(h)}$.

On choisit d'ordinaire une norme dans U_h de façon que pour le pas h tendant vers zéro elle change en une norme de fonctions données sur le segment tout entier, c'est-à-dire qu'on ait l'égalité

$$\lim_{h\to 0}\|[u]_h\|_{U_h} = \|u\|_U, \tag{4}$$

où $\|\cdot\|_U$ est la norme dans l'espace de fonctions sur le segment dont fait partie la solution $u(x)$. La norme

$$\|z^{(h)}\|_{U_h} = \max_m |z_m^{(h)}|$$

répond à cette condition si on prend pour U un espace de fonctions continues tel que

$$\|u\|_U = \max_{0<x<1} |u(x)|,$$

et si la fonction $[u]_h$ se confond par définition avec $u(x)$ aux points du réseau.

La norme

$$\|z^{(h)}\|_{U_h} = \sqrt{h\sum_m |z_m^{(h)}|^2} \tag{5}$$

est elle aussi une norme raisonnable. Elle vérifie la condition (4) si on prend pour U un espace de fonctions continues de norme

$$\|u\|_U = \sqrt{\int_0^1 u^2(x)\,dx},$$

la fonction $[u]_h$ étant toujours définie comme se confondant avec $u(x)$ aux points du réseau.

En cas de $u(x)$ discontinue mais de carré intégrable, U peut être un espace de fonctions de carré intégrable muni de la norme

$$||u||_U = \sqrt{\int_0^1 u^2(x)\,dx},$$

et la valeur u_n de $[u]_h$ se définit non pas d'après la formule $u_n = u(nh)$ qui est éventuellement dépourvue de sens, mais par

$$u_n = \frac{1}{h}\int_{x_n-h/2}^{x_n+h/2} u(x)\,dx.$$

On a alors pour la fonction discontinue $u(x)$

$$\lim_{h\to 0} ||[u]_h||_{U_h} = \sqrt{\int_0^1 u^2\,dx}.$$

Il est clair que la convergence

$$\lim_{h\to 0} ||[u]_h - u^{(h)}||_{U_h} = 0$$

au sens de la norme (1), i.e. la convergence uniforme, entraîne la convergence au sens de la norme (5), i.e. la convergence en moyenne, mais que la convergence en moyenne n'implique pas celle uniforme. De toutes les normes raisonnables vérifiant la condition (4) on choisit donc celle pour laquelle on démontre la convergence d'un schéma aux différences concret. Ce choix n'obéit à aucune règle universelle.

S'agissant d'équations différentielles ordinaires et d'équations aux différences associées que nous sommes en train d'étudier dans ce chapitre, on se contente d'ordinaire des normes (1), (5) ou d'une norme de la forme

$$||z^{(h)}||_{U_h} = \max\left[\max_m |z_m^{(h)}|,\quad \max_m \frac{|z_{m+1}^{(h)} - z_m^{(h)}|}{h}\right], \tag{6}$$

qui tient compte de la vitesse de variation de la fonction discrète lorsqu'on passe d'un point du réseau à l'autre. L'égalité (4) est juste pour cette norme si on prend pour U un espace de fonctions continûment dérivables de norme

$$||u||_U = \max\left[\max_{0<x<1} |u(x)|,\ \max_{0<x<1} |u'(x)|\right].$$

Dans le cas d'équations aux dérivées partielles et de schémas aux différences correspondants il y a parfois intérêt à introduire des normes assez ingénieuses adaptées à tel ou tel problème concret.

Passons au choix d'une norme dans l'espace F_h auquel appartient le second membre du schéma aux différences $L_h u^{(h)} = f^{(h)}$. Soulignons que la convergence $||[u]_h - u^{(h)}||_{U_h} \to 0$ du schéma aux différences pour la norme $||\cdot||_{U_h}$ choisie ne dépend pas de la façon dont on a introduit la norme $||\cdot||_{F_h}$ ni en général du fait de son existence. Si nous considérons F_h comme un espace vectoriel normé, c'est uniquement pour ramener la

démonstration de la convergence et la vérification de l'ordre de précision du schéma aux différences à la vérification de l'approximation d'un certain ordre et à celle de la stabilité.

La discussion du choix d'une norme dans F_h se fera sous l'hypothèse que le schéma aux différences $L_h u^{(h)} = f^{(h)}$ soit linéaire. Le seul but de cet artifice est d'éviter des raisonnements superflus.

Supposons que pour un choix fixe de la norme $\|\cdot\|^{(1)}$ le schéma aux différences $L_h u^{(h)} = f^{(h)}$ approche le problème $Lu = f$ pour la solution u à un ordre h^k et est stable. En vertu du théorème de convergence le schéma $L_h u^{(h)} = f^{(h)}$ est alors convergent et exact à l'ordre h^k :

$$\|[u]_h - u^{(h)}\|_{U_h} < Ch^k. \tag{7}$$

L'approximation signifie, rappelons-le, la validité d'une inégalité de la forme

$$\|L_h[u]_h - f^{(h)}\|^{(1)}_{F_h} \leqslant C_1 h^k. \tag{8}$$

La stabilité équivaut à dire que le problème $L_h u^{(h)} = f^{(h)}$ possède une solution unique quel que soit $f^{(h)} \in F_h$, de plus

$$\|u^{(h)}\|_{U_h} \leqslant C_2 \|f^{(h)}\|^{(1)}_{F_h}. \tag{9}$$

Si l'on choisit une autre norme $\|\cdot\|_{F_h}$ en posant

$$\|f^{(h)}\|^{(2)}_{F_h} = h\,\|f^{(h)}\|^{(1)}_{F_h}, \tag{10}$$

les inégalités (8) et (9) sont évidemment remplacées respectivement par

$$\left.\begin{aligned} &\|L_h[u]_h - f^{(h)}\|^{(2)}_{F_h} \leqslant Ch^{k+1}, \\ &\|u^{(h)}\|_{U_h} \leqslant \frac{C}{h}\,\|f^{(h)}\|^{(2)}_{F_h}. \end{aligned}\right\} \tag{11}$$

L'approximation sera donc non pas d'ordre k en h, mais d'ordre $k+1$. Cela ferait conclure à tort que l'ordre de précision du schéma aux différences est h^{k+1} et non pas h^k. C'est que l'inégalité (9) ne signifie plus la stabilité qui n'a en général pas lieu pour ce nouveau choix de la norme.

Si, à la place de (10), nous avions introduit une norme $\|\cdot\|^{(2)}_{F_h}$ définie par

$$\|f^{(h)}\|^{(2)}_{F_h} = \frac{1}{h}\,\|f^{(h)}\|^{(1)}_{F_h},$$

nous aurions

$$\|L_h[u]_h - f^{(h)}\|^{(2)}_{F_h} \leqslant C_1 h^{k-1}, \tag{12}$$

$$\|u^{(h)}\|_{U_h} \leqslant C_2 h\,\|f^{(h)}\|^{(2)}_{F_h} \tag{13}$$

au lieu de (8) et (9). L'inégalité (13) garantit la stabilité puisque $C_2 h$ ne peut être remplacé par une constante C_2 indépendante de h qu'en renforçant l'inégalité. L'inégalité (12) signifie une approximation d'ordre $k-1$ en h.

Pour la norme $\|\cdot\|^{(2)}_{F_h}$ ainsi choisie, nous ne pourrions garantir, en vertu du théorème de convergence, que l'ordre de précision $k-1$ du schéma aux différences $L_h u^{(h)} = f^{(h)}$, soit un ordre d'une unité inférieur à celui garanti par l'inégalité (7). La perte d'information sur l'ordre de précision est due à un choix malheureux de la norme dans l'espace F_h.

Pour mettre en évidence de façon correcte l'ordre de précision du schéma aux différences, on doit choisir une norme $\|\cdot\|_{F_h}$ telle que l'ordre d'approximation soit le plus élevé possible et la stabilité se conserve. Il n'existe pas de règle générale présidant à un tel choix de $\|\cdot\|_{F_h}$ (*). Il y a plus. Il n'est pas toujours possible de choisir une norme de façon qu'il y ait approximation et stabilité à la fois. Dans le cas contraire, en dépit de l'exemple du § 9, tout schéma aux différences serait convergent.

Nous allons cependant présenter une considération d'ordre général qui aide à choisir correctement une norme dans l'espace vectoriel F_h. En choisissant $\|\cdot\|_{F_h}$ on utilisera le fait que la solution du problème aux limites différentiel $Lu = f$ à la base du schéma aux différences $L_h u^{(h)} = f^{(h)}$ dépend de façon continue du second membre f.

Soit par exemple le problème

$$\frac{du}{dx} + Au = \varphi(x), \quad u(0) = a, \quad 0 \leqslant x \leqslant 1.$$

Les variations $\delta\varphi(x)$ et δa sur les seconds membres de l'équation et de la condition aux limites respectivement entraînent une variation $\delta u(x)$ du même ordre de grandeur sur la solution $u(x)$.

Considérons le schéma aux différences

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1}-u_n}{h} + Au_n = \varphi(x_n), & n = 0, 1, \ldots, N-1, \\ u_0 = a, \end{cases}$$

de sorte que

$$f^{(h)} = \begin{cases} \varphi(x_n), & n = 0, 1, \ldots, N-1, \\ a. \end{cases}$$

Définissons comme d'ordinaire une norme dans U_h par l'égalité

$$\|u^{(h)}\|_{U_h} = \max_m |u_m^{(h)}|.$$

On ne s'attend à la stabilité que si la norme

$$\|f^{(h)}\|_{F_h} = \left\| \begin{matrix} \varphi(x_n) \\ a \end{matrix} \right\|_{F_h}$$

dépend essentiellement de $\varphi(x_n)$ et de a. Elle peut s'écrire par exemple

$$\|f^{(h)}\|_{F_h} = \max\left[|a|, \max_m |\varphi_m|\right]. \tag{14}$$

(*) Cela est également vrai pour les schémas aux différences servant à résoudre les équations aux dérivées partielles.

En considérant plus haut (§12) un problème non linéaire plus général nous avons démontré la stabilité pour cette norme.

Il n'y aura pas de stabilité si la norme est choisie disons par la formule

$$\|f^{(h)}\|_{F_h} = \max\left[h\,|a|,\ \max_m |\varphi_m|\right],$$

où a figure avec un poids diminuant avec h.

La stabilité au sens de cette norme signifierait que la solution $u^{(h)}$ dépend de a plus faiblement que la solution u de l'équation différentielle. Or, pour h faible, la solution de l'équation aux différences diffère peu, vu la convergence (il y aurait convergence en cas de stabilité puisqu'il y a également approximation), de celle de l'équation différentielle et, la valeur initiale de a variant, elle doit varier à peu près comme la solution $u(x)$.

Soyons plus nets : pour ce choix de la norme le problème

$$\left.\begin{aligned} \frac{u_{n+1}-u_n}{h} + Au_n = \varphi_n, \quad n = 0, 1, \ldots, N-1, \\ u_0 = 0 \end{aligned}\right\}$$

approche le problème

$$\frac{du}{dx} + Au = \varphi(x), \quad u(0) = a$$

pour la solution $u(x)$ quel que soit a. S'il y a stabilité, la fonction $u^{(h)}$ indépendante de a devrait donc converger vers la solution $u(x)$ quel que soit a donné. Mais $u^{(h)}$ ne peut converger vers divers $u(x)$ à la fois.

Dans le cas du schéma aux différences

$$\left.\begin{aligned} \frac{u_{n+1}-2u_n+u_{n-1}}{h^2} + A\,\frac{u_{n+1}-u_{n-1}}{2h} + Bu_n = \varphi_n, \quad n = 1, \ldots, N, \\ u_0 = a, \\ \frac{u_1-u_0}{h} = b \end{aligned}\right\} \tag{15}$$

pour le problème

$$\left.\begin{aligned} \frac{d^2u}{dx^2} + A\,\frac{du}{dx} + Bu = \varphi(x), \\ u(0) = a, \\ \frac{du(0)}{dx} = b \end{aligned}\right\}$$

on établit, en se servant des mêmes considérations, que la norme

$$\|f^{(h)}\|_{F_h} = \left\| \begin{matrix} \varphi(x_m) \\ a \\ b \end{matrix} \right\|_{F_h}$$

dépend essentiellement de φ, a et b. Elle peut s'écrire

$$\|f^{(h)}\|_{F_h} = \max\left[|a|,\ |b|,\ \max_m |\varphi_m|\right], \tag{16}$$

mais on s'attendra en vain à la stabilité si l'on choisit pour norme $\|f^{(h)}\|_{F_h}$ par exemple la quantité

$$\|f^{(h)}\|_{F_h} = \max\left[|a|,\ h\,|b|,\ \max_m |\varphi_m|\right].$$

Donnons au schéma (15) un aspect légèrement différent :

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1}-2u_n+u_{n-1}}{h^2} + A\,\dfrac{u_{n+1}-u_{n-1}}{2h} + Bu_n = \varphi(x_n), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad u_0 = a, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad u_1 = a+bh, \end{cases} \tag{17}$$

de sorte que

$$f^{(h)} = \begin{cases} \varphi(x_n), & n = 1, 2, \ldots, N-1, \\ a, \\ a+bh. \end{cases}$$

On introduit une norme dans F_h en la définissant pour un élément arbitraire $g^{(h)} = \begin{cases} \psi_m, \\ \alpha, \\ \beta \end{cases}$ par une formule du type

$$\|g^{(h)}\|_{F_h} = \max\left[|\alpha|,\ \frac{|\beta-\alpha|}{h},\ \max_m |\psi_m|\right], \tag{18}$$

où $|\beta-\alpha|$ figure avec un poids $1/h$ croissant pour $h \to 0$. En effet, une variation h sur α ou β équivaut à faire varier u_0 ou u_1 de la même quantité, $\dfrac{|u_1-u_0|}{h}$ variant alors d'une quantité de l'ordre de 1. Pour un schéma stable ce dernier fait entraînera une variation de l'ordre de 1 sur la solution de l'équation

$$\frac{u_{n+1}-2u_n+n_{n-1}}{h^2} + A\,\frac{u_{n+1}-u_{n-1}}{2h} + Bu_n = \varphi_n,$$

puisque si le rapport $\dfrac{|u_1-u_0|}{h}$ varie de $O(1)$, cela équivaut à une variation de l'ordre de 1 sur le second membre de la condition $\dfrac{du(0)}{dx} = b$ du problème différentiel.

Il n'y aura pas de stabilité pour une norme définie par la formule

$$\|g^{(h)}\|_{F_h} = \max\left[|\alpha|,\ |\beta|,\ \max_m |\psi_m|\right],$$

i.e. comme elle l'a été plus haut quand nous avons utilisé l'espace F_h pour équiper le schéma aux différences (15). L'ordre d'approximation des sché-

mas (15) et (17) pour les normes respectives (16) et (18) est le premier en h pour les deux. La stabilité de ces schémas pour les normes (16) et (18) sera démontrée au §14.

§14. Critère suffisant de stabilité des schémas aux différences pour le problème de Cauchy

Nous allons étudier la stabilité des schémas aux différences $L_h u^{(h)} = f^{(h)}$ pour la résolution d'un problème différentiel avec conditions initiales (problème de Cauchy) et nous examinerons à cet effet des exemples caractéristiques de schémas aux différences approchant les problèmes

$$Lu = \begin{cases} \dfrac{du}{dx} + Au = \varphi(x), & 0 \leqslant x \leqslant 1, \\ u(0) = a, \end{cases} \tag{1}$$

$$Lu \equiv L\begin{pmatrix} v \\ w \end{pmatrix} = \begin{cases} \dfrac{dv}{dx} + Av + Bw = p(x), & 0 \leqslant x \leqslant 1, \\ \dfrac{dw}{dx} + Cv + Dw = q(x), & 0 \leqslant x \leqslant 1, \\ v(0) = a, \\ w(0) = b, \end{cases} \tag{2}$$

$$Lu \equiv \begin{cases} \dfrac{d^2u}{dx^2} + A\dfrac{du}{dx} + Bu = \varphi(x), & 0 \leqslant x \leqslant 1, \\ u(0) = a, \\ \dfrac{du(0)}{dx} = b. \end{cases} \tag{3}$$

Pour que la notion de stabilité du schéma $L_h u^{(h)} = f^{(h)}$ ait un sens, il faut définir les espaces vectoriels normés U_h et F_h dont font partie respectivement la table $[u]_h$ de la solution u du problème différentiel, table qu'on se propose de calculer approximativement, et le second membre $f^{(h)}$ du schéma aux différences.

Rappelons que le schéma aux différences $L_h u^{(h)} = f^{(h)}$ d'opérateur linéaire L_h est dit *stable* si le problème $L_h u^{(h)} = f^{(h)}$ admet une solution unique $u^{(h)} \in U_h$ quel que soit $f^{(h)} \in F_h$ et si, de plus,

$$\|u^{(h)}\|_{U_h} \leqslant C\,\|f^{(h)}\|.$$

En résolvant le problème de Cauchy on calcule d'ordinaire $u^{(h)}$ de proche en proche en passant d'un point du réseau aux différences au point voisin. A condition d'obtenir à chaque pas du processus de calcul une estimation de la croissance de la solution $u^{(h)} = \{u_n^{(h)}\}$, on a là un procédé d'étude de la stabilité parmi les plus universels. Voici son idée.

1. Exemple préliminaire. Prenons pour commencer le schéma aux diffé-

rences simple qui nous est déjà familier

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{n+1}-u_n}{h} + Au_n = \varphi_n, \quad n = 0, 1, \ldots, N-1, \\ \qquad\qquad (h = 1/N), \\ \qquad\qquad u_0 = a \end{cases} \tag{4}$$

pour le problème (1). Il s'écrit sous forme récurrentielle

$$\left.\begin{aligned} u_{n+1} &= (1-Ah)u_n + h\varphi_n, \quad n = 0, 1, \ldots, N-1, \\ u_0 &= a. \end{aligned}\right\} \tag{5}$$

D'où

$$\left.\begin{aligned} u_1 &= (1-Ah)u_0 + h\varphi_0, \\ u_2 &= (1-Ah)^2 u_0 + h[(1-Ah)\varphi_0 + \varphi_1], \\ u_3 &= (1-Ah)^3 u_0 + h[(1-Ah)^2\varphi_0 + (1-Ah)\varphi_1 + \varphi_2], \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ u_n &= (1-Ah)^n u_0 + h[(1-Ah)^{n-1}\varphi_0 + (1-Ah)^{n-2}\varphi_1 + \ldots + \varphi_{n-1}]. \end{aligned}\right\} \tag{6}$$

Définissons les normes dans U_h et F_h par les égalités

$$\|u^{(h)}\|_{U_h} = \max_{0 \leqslant n \leqslant N} |u_n^{(h)}|, \tag{7}$$

$$\|f^{(h)}\|_{F_h} = \left\| \begin{matrix} \varphi_m \\ a \end{matrix} \right\|_{F_h} = \max \left[|a|, \max_{0 \leqslant m \leqslant N} |\varphi_m| \right]. \tag{8}$$

Utilisons le fait que l'expression $(1-Ah)^n$ est bornée pour $n \leqslant N = 1/h$:

$$|(1-Ah)^n| < C_1. \tag{9}$$

Moyennant (9) nous déduisons de la formule (6)

$$\begin{aligned} |u_n| &\leqslant C_1 |u_0| + hNC_1 \max_m |\varphi_m| = \\ &= C_1 |a| + C_1 \max_m |\varphi_m| \leqslant 2C_1 \|f^{(h)}\|_{F_h}. \end{aligned} \tag{10}$$

$n = 0, 1, \ldots, N$ étant quelconque, il en résulte

$$\|u^{(h)}\|_{U_h} \leqslant 2C_1 \|f^{(h)}\|_{F_h}, \tag{11}$$

ce qui démontre la stabilité.

2. Schéma aux différences sous forme canonique. Introduisons de nouvelles notations en posant

$$u_n = y_n, \quad R_h = (1-Ah), \quad \varrho_n = \varphi_n. \tag{12}$$

Les égalités (5) se mettent alors sous la forme

$$\left.\begin{aligned} y_{n+1} &= R_h y_n + h\varrho_n, \\ y_0 &\text{ donné.} \end{aligned}\right\} \tag{13}$$

Répétons tous les calculs précédents avec les notations (12). Les égalités (6) prennent alors la forme

$$\left.\begin{aligned} y_1 &= R_h y_0 + h\varrho_0, \\ y_2 &= R_h^2 y_0 + h[R_h\varrho_0 + \varrho_1], \\ y_3 &= R_h^3 y_0 + h[R_h^2\varrho_0 + R_h\varrho_1 + \varrho_2], \\ &\dots\dots\dots\dots\dots\dots\dots\dots \\ y_n &= R_h^n y_0 + h[R_h^{n-1}\varrho_0 + R_h^{n-2}\varrho_1 + \dots + \varrho_{n-1}]. \end{aligned}\right\} \tag{6'}$$

D'où

$$\max_n |y_n| \leqslant \max_n |R_h^n| \cdot \left[|y_0| + hN \max_n |\varrho_n|\right].$$

Les normes $\|\cdot\|_{U_h}$ et $\|\cdot\|_{F_h}$ sont maintenant définies par

$$\|u^{(h)}\|_{U_h} = \max_n |y_n|, \tag{7'}$$

$$\|f^{(h)}\|_{F_h} = \max\left[|y_0|, \max_n |\varrho_n|\right]. \tag{8'}$$

Compte tenu de $Nh = 1$, nous pouvons écrire

$$\|u^{(h)}\|_{U_h} \leqslant \max_n |R_h^n| \cdot 2\,\|f^{(h)}\|_{F_h}.$$

La démonstration de la stabilité sera achevée si nous établissons que l'ensemble de nombres $|R_h^n|$ est borné uniformément en h, c.-à-d. que

$$|R_h^n| \leqslant C, \quad n = 1, 2, \dots, N. \tag{9'}$$

Mais

$$|R_h^n| \leqslant (1 - Ah)^n \leqslant (1 - Ah)^N \leqslant e^{|A|} = C.$$

L'écriture (13) du schéma aux différences a permis de ramener la démonstration de la stabilité à l'évaluation de $|R_h^n|$, ce qui constitue un procédé très commode. Si bien que nous procéderons de même avec tous les autres schémas aux différences pour la résolution des problèmes avec conditions initiales, à savoir nous les ramènerons à la forme canonique (13) pour y_n, ϱ_n et R_h différant d'un cas à l'autre.

Mettons par exemple sous la forme (13) le schéma aux différences

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{v_{n+1} - v_n}{h} + Av_n + Bw_n = p_n, & n = 0, 1, \dots, N-1, \\ \dfrac{w_{n+1} - w_n}{h} + Cv_n + Dw_n = q_n, & n = 0, 1, \dots, N-1, \\ v_0 = a, \\ w_0 = b \end{cases} \tag{14}$$

approchant le problème de Cauchy (2) pour un système d'équations différen-

tielles. Ici

$$f^{(h)} = \begin{cases} p_n, & n = 0, 1, \ldots, N-1, \\ q_n, & n = 0, 1, \ldots, N-1, \\ a, & \\ b. & \end{cases}$$

Ecrivons le schéma aux différences (14) comme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{\begin{bmatrix} v_{n+1} \\ w_{n+1} \end{bmatrix} - \begin{bmatrix} v_n \\ w_n \end{bmatrix}}{h} + \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{bmatrix} v_n \\ w_n \end{bmatrix} = \begin{bmatrix} p_n \\ q_n \end{bmatrix}, \quad n = 0, 1, \ldots, N-1, \\ \begin{bmatrix} v_0 \\ w_0 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}, \end{cases}$$

$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ étant une matrice du second ordre. Mettons cette équation aux différences vectorielle sous forme d'une relation récurrentielle :

$$\left.\begin{aligned} \begin{bmatrix} v_{n+1} \\ w_{n+1} \end{bmatrix} &= \begin{pmatrix} 1-Ah & -Bh \\ -Ch & 1-Dh \end{pmatrix} \begin{bmatrix} v_n \\ w_n \end{bmatrix} + h \begin{bmatrix} p_n \\ q_n \end{bmatrix}, \\ \begin{bmatrix} v_0 \\ w_0 \end{bmatrix} &= \begin{bmatrix} a \\ b \end{bmatrix}. \end{aligned}\right\}$$

Si l'on pose

$$y_n = u_n = \begin{bmatrix} v_n \\ w_n \end{bmatrix}, \quad R_h = \begin{pmatrix} 1-Ah & -Bh \\ -Ch & 1-Dh \end{pmatrix}, \quad \varrho_n = \begin{bmatrix} p_n \\ q_n \end{bmatrix},$$

la dernière écriture prend la forme cherchée (13).

3. Stabilité en tant que borne des normes des puissances de R_h. L'écriture (13) entraîne (6'). Cela est vrai pour toutes les équations de la forme (13) quelles que soient la dimension de l'espace vectoriel Y dont font partie les vecteurs y_n et ϱ_n et la forme de l'opérateur linéaire R_h.

Si l'on munit l'espace Y d'une norme $\|\cdot\|_Y$, les égalités (6') entraînent l'estimation

$$\|y_n\|_Y \leqslant \|R_h^n\|_Y \cdot \|y_0\|_Y + h[\|R_h^{n-1}\|_Y \cdot \|\varrho_0\|_Y + \ldots + \|\varrho_{n-1}\|_Y]. \quad (15)$$

Rappelons qu'on appelle *norme* $\|T\|$ d'un opérateur linéaire T d'un espace vectoriel normé Y dans lui-même le nombre

$$\|T\| = \sup_{x \in Y} \frac{\|Tx\|}{\|x\|} \quad \text{ou} \quad \|T\| = \sup_{\|x\|=1,\, x \in Y} \|Tx\|.$$

Il en résulte, vu les propriétés d'une norme vectorielle,

$$\|Tx\| \leqslant \|T\| \cdot \|x\|,$$
$$\|\lambda T\| = |\lambda| \cdot \|T\|, \quad \lambda \text{ étant quelconque},$$
$$\|T^m\| \leqslant \|T\|^m.$$

Nous nous sommes servi des deux premières propriétés pour obtenir l'estimation (15).

Cette dernière entraîne évidemment

$$\max_{0<n<N} \|y_n\|_Y \leqslant \max_{0<n<N} \|R_h^n\|_Y \left[\|y_0\|_Y + Nh \max_{0<n<N} \|\varrho_n\|_Y\right]. \tag{16}$$

Supposons que le schéma aux différences $L_h u^{(h)} = f^{(h)}$ est réduit à la forme canonique (13) *et que les normes sur les espaces U_h, F_h et Y sont choisies de façon qu'on ait les inégalités*

$$\left.\begin{aligned} \|u^{(h)}\|_{U_h} &\leqslant C_2 \max_{0<n<N} \|y_n\|_Y, \\ \|y_0\|_Y &\leqslant C_2 \|f^{(h)}\|_{F_h}, \\ \|\varrho_n\|_Y &\leqslant C_2 \|f^{(h)}\|_{F_h}. \end{aligned}\right\} \tag{17}$$

Pour qu'il y ait stabilité

$$\|u^{(h)}\|_{U_h} \leqslant C \|f^{(h)}\|_{F_h}, \tag{18}$$

il suffit alors que les normes $\|R_h^m\|_Y$ des puissances des opérateurs R_h soient bornées uniformément en h, c'est-à-dire qu'on ait la condition

$$\|R_h^n\|_Y \leqslant C_3, \quad n = 1, 2, \ldots, N.$$

Ceci étant, on peut prendre pour C figurant dans la définition de la stabilité (18) le nombre

$$C = 2C_2^2 C_3.$$

Cette assertion se démontre moyennant la chaîne suivante d'inégalités évidentes écrites compte tenu de $Nh = 1$ et des conditions (17) et (18) :

$$\|u^{(h)}\|_{U_h} \leqslant C_2 \max_n \|y_n\|_Y \leqslant C_2 \max_n \|R_h^n\| \,[C_2 + \\ + C_2] \|f^{(h)}\|_{F_h} \leqslant C_2 C_3 [C_2 + C_2] \|f^{(h)}\|_{F_h},$$

ou

$$\|u^{(h)}\|_{U_h} \leqslant C \|f^{(h)}\|_{F_h}.$$

4. Exemples d'étude de la stabilité.

Exemple 1. Passons à l'analyse de la stabilité du schéma aux différences (14) pour un système d'équations différentielles. Munissons U_h et F_h de normes définies par les égalités

$$\|u^{(h)}\|_{U_h} = \left\| \begin{Bmatrix} v_n \\ w_n \end{Bmatrix} \right\|_{U_h} = \max\left[\max_n |v_n|, \max_n |w_n|\right],$$

$$\|f^{(h)}\|_{F_h} = \left\| \begin{Bmatrix} p_n \\ q_n \\ a \\ b \end{Bmatrix} \right\|_{F_h} = \max\left[|a|, |b|, \max_n |p_n|, \max_n |q_n|\right].$$

En introduisant les notations

$$y_n = \begin{bmatrix} v_n \\ w_n \end{bmatrix}, \quad R_h = \begin{pmatrix} 1-Ah & -Bh \\ -Ch & 1-Dh \end{pmatrix}, \quad \varrho_n = \begin{bmatrix} p_n \\ q_n \end{bmatrix}, \quad y_0 = \begin{bmatrix} a \\ b \end{bmatrix}$$

le système considéré d'équations aux différences se ramène, nous l'avons vu, à la forme canonique (13).

On prend une norme dans l'espace bidimensionnel Y dont font partie y_n et ϱ_n en posant

$$||y_n||_Y = \left\| \begin{bmatrix} y_n^{(1)} \\ y_n^{(2)} \end{bmatrix} \right\|_Y = \max\,[|y_n^{(1)}|, |y_n^{(2)}|].$$

Les normes dans U_h, F_h et Y vérifient les conditions (17). Il suffit donc de démontrer que

$$||R_h^n||_Y \leqslant M, \quad n = 1, 2, \ldots, N, \quad M = \text{const},$$

pour qu'il y ait stabilité.

Notons que pour la norme vectorielle choisie dans Y, la norme de tout opérateur linéaire

$$T = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix}$$

est donnée par la formule

$$||T|| = \max\,[|t_{11}|+|t_{12}|, |t_{21}|+|t_{22}|], \tag{19}$$

étant donné que $\max\limits_{||x||=1} ||Tx||_Y = ||T||_Y$ est atteint pour au moins un des vecteurs $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $x = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

En vertu de la formule (19) donnant $||T||$ on a

$$||R_h||_Y = \left\| \begin{pmatrix} 1-Ah & -Bh \\ -Ch & 1-Dh \end{pmatrix} \right\|_Y \leqslant$$
$$\leqslant \max\,[|1-Ah|+|Bh|, |1-Dh|+|Ch|] = 1+C'h.$$

Donc

$$||R_h^n||_Y \leqslant ||R_h||_Y^n \leqslant (1+C'h)^N \leqslant e^{C'} = M, \quad n = 1, 2, \ldots, N,$$

et la stabilité est démontrée.

Exemple 2. Soit le schéma

$$\left.\begin{aligned} \frac{u_{n+1}-u_{n-1}}{2h} + Au_n &= \varphi_n, \quad n = 1, 2, \ldots, N-1, \\ u_0 &= \alpha, \\ u_1 &= \beta, \end{aligned}\right\} \tag{20}$$

qui pour $\alpha = a$, $\beta = (1-Ah)\,a+h\varphi_0$ approche le problème de Cauchy (1) à l'ordre 2 en h.

Introduisons les normes $\|\cdot\|_{U_h}$ et $\|\cdot\|_{F_h}$ par les égalités

$$\|u^{(h)}\|_{U_h} = \max_n |u_n|,$$

$$\|f^{(h)}\|_{F_h} = \left\| \begin{bmatrix} \varphi_n \\ \alpha \\ \beta \end{bmatrix} \right\|_{F_h} = \max \left[|\alpha|,\ |\beta|,\ \max_n |\varphi_n| \right].$$

Pour établir la stabilité nous essayerons de mettre le schéma aux différences sous forme (13) et de ramener la démonstration à l'obtention de l'estimation $\|R_h^n\|_Y \leqslant C$. Récrivons l'équation aux différences (20) comme

$$u_{n+1} = u_{n-1} - 2Ahu_n + 2h\varphi_n.$$

L'équation reliant non pas deux mais trois valeurs successives u_{n-1}, u_n, u_{n+1}, le passage à la forme (13) s'en trouve entravé. On tourne cette difficulté en posant

$$y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}.$$

Le couple d'égalités

$$\left.\begin{aligned} u_{n+1} &= u_{n-1} - 2Ahu_n + 2h\varphi_n, \\ u_n &= u_n \end{aligned}\right\} \tag{21}$$

exprime alors les composantes du vecteur y_n en fonction de celles du vecteur y_{n-1} :

$$\begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix} = \begin{pmatrix} -2Ah & 1 \\ 1 & 0 \end{pmatrix} \begin{bmatrix} u_n \\ u_{n-1} \end{bmatrix} + h \begin{bmatrix} 2\varphi_n \\ 0 \end{bmatrix}.$$

Ainsi, nous avons réduit le problème (20) à la forme (13), où

$$R_h = \begin{pmatrix} -2Ah & 1 \\ 1 & 0 \end{pmatrix}, \quad \varrho_n = \begin{bmatrix} 2\varphi_n \\ 0 \end{bmatrix},$$

$$y_0 = \begin{bmatrix} (1-Ah)a + h\varphi_0 \\ a \end{bmatrix}.$$

Munissons l'espace bidimensionnel Y auquel appartiennent y_n et ϱ_n d'une norme définie par

$$\left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_Y = \max(|\alpha|,\ |\beta|).$$

On constate alors aisément que les normes

$$\|u^{(h)}\|_{U_h}, \quad \|f^{(h)}\|_{F_h}, \quad \|\varrho_n\|_Y, \quad \|y_0\|_Y$$

satisfont aux conditions (17). L'estimation

$$\|R_h^n\|_Y \leqslant \|R_h\|_Y^n = \left\| \begin{pmatrix} -2Ah & 1 \\ 1 & 0 \end{pmatrix} \right\|_Y^n \leqslant$$

$$\leqslant (1+2|Ah|)^n \leqslant e^{2|A|}, \quad n = 1, 2, \ldots, N,$$

démontre donc la stabilité.

Exemple 3. Etudions la stabilité du schéma aux différences

$$\left.\begin{aligned} \frac{u_{n+1}-2u_n+u_{n-1}}{h^2} + A\frac{u_{n+1}-u_{n-1}}{2h} + Bu_n &= \varphi_n, \quad n = 1, 2, \ldots, N-1, \\ u_0 &= a, \\ \frac{u_1-u_0}{h} &= b, \end{aligned}\right\} \tag{22}$$

qui approche pour un choix usuel des normes le problème de Cauchy (3). Définissons deux normes $\|u^{(h)}\|_{U_h}$ et $\|f^{(h)}\|_{F_h}$ par les égalités

$$\left.\begin{aligned} \|u^{(h)}\|_{U_h} &= \max_n |u_n|, \\ \|f^{(h)}\|_{F_h} &= \left\| \begin{matrix} \varphi_n \\ a \\ b \end{matrix} \right\|_{F_h} = \max\left[|a|, |b|, \max_n |\varphi_n|\right]. \end{aligned}\right\} \tag{23}$$

On ramène le schéma en question à la forme canonique (13) en posant, comme dans l'exemple 2,

$$y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}. \tag{24}$$

En vertu de l'équation aux différences donnée, les composantes du vecteur $y_n = \begin{bmatrix} y_n^{(1)} \\ y_n^{(2)} \end{bmatrix}$ s'expriment alors de façon unique en fonction de celles du vecteur y_{n-1} par les formules

$$\left.\begin{aligned} y_{n+1}^{(1)} &= \frac{2}{2+Ah}\left[(2-Bh^2)y_n^{(1)} - \frac{2-Ah}{2}y_n^{(2)} + h^2\varphi_{n+1}\right], \\ y_{n+1}^{(2)} &= y_n^{(1)}. \end{aligned}\right\} \tag{25}$$

Donc

$$y_{n+1} = R_h y_n + h\varrho_n, \quad n = 0, 1, \ldots, N-2,$$

où

$$R_h = \begin{pmatrix} \frac{4-2Bh^2}{2+Ah} & -\frac{2-Ah}{2+Ah} \\ 1 & 0 \end{pmatrix}, \quad \varrho_n = \begin{bmatrix} \frac{2h^2}{2+Ah}\varphi_{n+1} \\ 0 \end{bmatrix}. \tag{26}$$

Calculons le vecteur y_0 à l'aide des conditions $u_0 = a$, $\frac{u_1 - u_0}{h} = b$ (voir (22)), il vient

$$y_0 = \begin{bmatrix} a+bh \\ a \end{bmatrix}, \tag{27}$$

ce qui achève de réduire le schéma considéré à la forme (13).

On voit facilement qu'en définissant la norme du vecteur $\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ comme $\max(|\alpha|, |\beta|)$, dans le cas de notre opérateur R_h la démonstration de la stabilité n'est plus si facile parce que $\|R_h\| \approx 2$ et $\|R_h\|^n \to \infty$. Aussi la norme sur Y ne sera pas définie comme dans l'exemple 2. Nous posons notamment

$$\left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y_h} = \max\left[|\alpha|, \left|\frac{\beta - \alpha}{h}\right|\right].$$

L'indice h de Y doit souligner la dépendance de la norme par rapport à h. Pour le choix fait des normes, il y a entre $\|u^{(h)}\|_{U_h}$, $\|f^{(h)}\|_{F_h}$, $\|\varrho_n\|_{Y_h}$, $\|y_0\|_{L_h}$ les relations (17). Il reste à tester la condition $\|R_h^n\|_{Y_h} \leqslant C$, $n = 1, 2, \ldots, N$. Nous connaissons la formule (19) exprimant la norme d'un opérateur en fonction d'éléments de la matrice qui le définit si la norme dans Y est définie par

$$\left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y} = \max[|\alpha|, |\beta|].$$

Ramenons à ce cas le calcul de la norme d'un opérateur dans l'espace Y_h :

$$\left\| \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \right\|_{Y_h} = \left\| \begin{pmatrix} 1 & 0 \\ 1/h & -1/h \end{pmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y} = \left\| S \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y},$$

avec $S = \begin{pmatrix} 1 & 0 \\ 1/h & -1/h \end{pmatrix}$. Montrons que toute transformation linéaire T opérant dans Y vérifie l'égalité $\|T\|_{Y_h} = \|STS^{-1}\|_Y$. En effet,

$$\left\| T \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y_h} = \left\| ST \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_{Y} = \left\| STS^{-1}\left(S \begin{bmatrix} \alpha \\ \beta \end{bmatrix}\right) \right\|_{Y}.$$

Ensuite

$$\|T\|_{Y_h} = \max_x \frac{\|Tx\|_{Y_h}}{\|x\|_{Y_h}} = \max_{x \in Y} \frac{\|STS^{-1}Sx\|_Y}{\|Sx\|_Y} =$$

$$= \max_{v \in Y} \frac{\|STS^{-1}v\|_Y}{\|v\|_Y} = \|STS^{-1}\|_Y.$$

Notons à présent que

$$\|R_h^n\|_{Y_h} = \|SR_h^nS^{-1}\|_Y = \|(SR_hS^{-1})^n\|_Y \leqslant \|SR_hS^{-1}\|_Y^n.$$

Comme $S^{-1} = \begin{pmatrix} 1 & 0 \\ 1 & -h \end{pmatrix}$, on a

$$SR_hS^{-1} = \begin{pmatrix} 1-\dfrac{2B}{2+Ah}h & \dfrac{2-Ah}{2+Ah}h \\ -\dfrac{2B}{2+Ah}h & 1-\dfrac{2A}{2+Ah}h \end{pmatrix}.$$

Donc

$$\|SR_hS^{-1}\|_Y \leqslant 1+Ch,$$

C étant une constante indépendante de h choisie par la condition

$$1+Ch \geqslant \max\left[\left|1-\frac{2B}{2+Ah}h\right|+\left|\frac{2-Ah}{2+Ah}h\right|,\ \left|\frac{2B}{2+Ah}h\right|+\left|1-\frac{2A}{2+Ah}h\right|\right].$$

En particulier, pour h suffisamment petits, cette condition est évidemment vérifiée par le nombre $C = 1+2|A|+2|B|$.

Ainsi,

$$\|R_h^n\|_{Y_n} \leqslant \|SR_hS^{-1}\|_Y^n \leqslant (1+Ch)^N \leqslant e^C, \quad n = 1, 2, \ldots, N,$$

ce qui garantit la stabilité du schéma étudié.

5. Non-unicité de l'écriture canonique. Un schéma aux différences peut être ramené à la forme canonique (13) par diverses techniques. En posant $y'_n = Ty_n$ avec T une transformation linéaire quelconque de l'espace Y dont font partie y_n et ϱ_n, on passe à l'écriture

$$\left.\begin{aligned} &y'_{n+1} = R'_h y'_n + h\varrho'_n, \\ &y'_0 \text{ donné.} \end{aligned}\right\} \tag{13'}$$

Ici $R'_h = TR_hT^{-1}$, $\varrho'_n = T\varrho_n$, $y'_0 = Ty_0$.

En faisant $y_n = \begin{bmatrix} u_n \\ \dfrac{u_{n+1}-u_n}{h} \end{bmatrix}$ au lieu de $y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}$ dans l'exemple 3, nous aurions abouti à l'écriture (13) où

$$R_h = \begin{pmatrix} 1 & h \\ -\dfrac{2hB}{2+hA} & \dfrac{2-hA-2h^2B}{2+hA} \end{pmatrix}, \quad \varrho_n = \begin{bmatrix} 0 \\ \dfrac{2}{2+hA}\varphi_{n+1} \end{bmatrix}.$$

En choisissant la norme dans Y selon la formule $\left\|\begin{bmatrix}\alpha\\ \beta\end{bmatrix}\right\|_Y = \max(|\alpha|, |\beta|)$, les conditions (17) se seraient trouvées satisfaites. Le caractère borné de $\|R_h^n\|_Y$ est évident :

$$\|R_h^n\|_Y \leqslant \|R_h\|_Y^n \leqslant (1+Ch)^N < e^C,$$

avec C choisi à partir de la condition

$$1+Ch \geqslant \max\left(1+h,\ \left|\frac{2hB}{2+hA}\right|+\left|\frac{2-hA-2h^2B}{2+hA}\right|\right) =$$
$$= \max\left(1+h,\ 1+\frac{2(|A|+|B|h)}{2-|A|h}h\right).$$

Le choix de la dimension de Y comporte lui aussi un certain arbitraire. Nous aurions pu, par exemple, poser non pas $y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}$ mais $y_n = \begin{bmatrix} u_{n+2} \\ u_{n+1} \\ u_n \end{bmatrix}$, artifice qui n'aurait pas d'ailleurs simplifié l'étude de la stabilité dans cet exemple.

Faisons le bilan de nos raisonnements. Les exemples considérés permettent de conclure qu'en étudiant la stabilité du schéma aux différences $L_h u^{(h)} = f^{(h)}$ pour la résolution du problème de Cauchy à coefficients constants, il y a intérêt à réduire ce schéma à la forme (13) :

$$\left.\begin{aligned} y_{n+1} &= R_h y_n + h\varrho_n, \quad n = 0, 1, \ldots, \\ y_0 &\text{ donné.} \end{aligned}\right\}$$

Si l'espace dont font partie y_n et ϱ_n est muni d'une norme de façon à satisfaire aux conditions

$$\left.\begin{aligned} \|u^{(h)}\|_{U_h} &\leqslant C_2 \max_n \|y_n\|, \\ \|\varrho_n\| &\leqslant C_2 \|f^{(h)}\|_{F_h}, \\ \|y_0\| &\leqslant C_2 \|f^{(h)}\|_{F_h}, \end{aligned}\right\} \tag{28}$$

il suffit, pour avoir la stabilité, que les normes des puissances de l'opérateur R_h soient bornées uniformément en h

$$\|R_h^n\| \leqslant C_3, \quad n = 1, 2, \ldots, N.$$

La condition suffisante en est évidemment l'inégalité

$$\|R_h\| < 1 + C'h,$$

avec C' indépendant de h.

La constante C figurant dans la définition de la stabilité

$$\|u^{(h)}\|_{U_h} \leqslant C \|f^{(h)}\|_{F_h}$$

peut être dans ce cas prise sous la forme

$$C = 2C_2^2 C_3. \tag{29}$$

Exercices

1. Démontrer la stabilité des schémas aux différences suivants servant à résoudre le problème $u' + Au = \varphi(x)$, $u(0) = a$. Trouver la constante C figurant dans la définition de la stabilité $\|u^{(h)}\|_{U_h} \leqslant C \|f^{(h)}\|_{F_h}$.

$$\text{a)} \quad \left.\begin{aligned} \frac{u_{n+1} - u_n}{h} + A(nh)u_n &= \varphi_n, \quad n = 0, 1, 2, \ldots, N-1, \\ u_0 &= a \end{aligned}\right\}$$

si $|A(x)| \leqslant M = \text{const}$ et les normes sont introduites au moyen des égalités

$$\|u^{(h)}\|_{U_h} = \max_n |u_n|, \qquad \|f^{(h)}\|_{F_h} = \max\left[|a|, \max_n |\varphi_n|\right].$$

$$\text{b)} \quad \left.\begin{aligned} \frac{u_{n+1}-u_n}{h} + Au_{n+1} &= \varphi_n, \\ u_0 &= a. \end{aligned}\right\} \text{ Pour les normes voir a).}$$

$$\text{c)} \quad \left.\begin{aligned} \frac{u_{n+1}-u_n}{h} + A\,\frac{u_n+u_{n+1}}{2} &= \varphi\left[\left(n+\frac{1}{2}\right)h\right], \quad n = 0, 1, \ldots, N-1, \\ u_0 &= a, \end{aligned}\right\}$$

$$\|u^{(h)}\|_{U_h} = \max_n |u_n|, \qquad \|f^{(h)}\|_{F_h} = \max\left(|a|, \max_n \left|\varphi\left[\left(n+\frac{1}{2}\right)h\right]\right|\right).$$

2. Le même exercice sous l'hypothèse que $u_n = \begin{bmatrix} u_n^{(1)} \\ u_n^{(2)} \end{bmatrix}$ est un vecteur, $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ une matrice et $a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$, $\varphi_n = \begin{bmatrix} \varphi_n^{(1)} \\ \varphi_n^{(2)} \end{bmatrix}$ sont des vecteurs. Les normes sont données sous la forme

$$\|u^{(h)}\|_{U_h} = \max_n (|u_n^{(1)}| + |u_n^{(2)}|),$$

$$\|f^{(h)}\|_{F_h} = \max\left[|a_1| + |a_2|, \max_n (|\varphi_n^{(1)}| + |\varphi_n^{(2)}|)\right], \quad |A_{ij}(x)| \leqslant M.$$

3. Réduire à la forme canonique $y_{n+1} = R_h y_n + h\varrho_n$, y_0 donné, l'équation aux différences

$$\frac{u_{n+2} - 2u_{n+1} + 3u_n - 4u_{n-1}}{h} - 5u_n = \varphi_n, \qquad n = 1, 2, \ldots,$$

à condition de poser $y_n = \begin{bmatrix} u_{n+2} \\ u_{n+1} \\ u_n \end{bmatrix}$.

§15. Critère spectral nécessaire de stabilité

Nous avons montré au paragraphe précédent que le fait de ramener le schéma aux différences pour le problème de Cauchy à coefficients constants

$$L_h u^{(h)} = f^{(h)} \tag{1}$$

à la forme

$$\left.\begin{aligned} &y_{n+1} = R_h y_n + h\varrho_n, \quad n = 0, 1, \ldots, \\ &y_0 \text{ donné} \end{aligned}\right\} \tag{2}$$

peut être utilisé pour démontrer la stabilité : dans certaines conditions ((17) §14), l'estimation

$$\|R_h^n\|_Y \leqslant C, \quad n = 1, 2, \ldots, N, \tag{3}$$

est suffisante pour avoir la stabilité.

Nous allons maintenant montrer que dans certaines conditions naturelles elle est également nécessaire et établir que, quelle que soit la norme

choisie, la condition nécessaire de (3) est que le spectre de la matrice R_h, i.e. l'ensemble de racines de l'équation

$$\det(R_h - \lambda E) = 0, \tag{4}$$

se trouve dans le cercle

$$|\lambda| < 1 + Ch, \tag{5}$$

avec C indépendant de h.

1. Une condition nécessaire de stabilité est que les normes des puissances de l'opérateur R_h soient bornées. Nous constatons, en reprenant tous les procédés examinés de réduction des équations aux différences à la forme (2), que si leurs seconds membres sont nuls, alors $\varrho_n \equiv 0$.

Soient les constantes $M_1 = M_1(h) > 0$ et $M_2 = M_2(h) > 0$ telles qu'on ait les inégalités

$$\|u^{(h)}\|_{U_h} \geqslant M_1 \max_n \|y_n\|, \tag{6}$$

$$\|y_0\| \geqslant M_2 \|f^{(h)}\|_{F_h} \tag{7}$$

(la dernière dans la condition $\varrho_n \equiv 0$). Les seconds membres d'une équation aux différences (ou d'un système d'équations aux différences) étant nuls, l'équation (2) prend la forme

$$y_{n+1} = R_h y_n,$$

d'où

$$y_n = R_h^n y_0. \tag{8}$$

Ensuite, en vertu de (6) et (8),

$$\|u^{(h)}\|_{U_h} \geqslant M_1 \max_n \|R_h^n y_0\|. \tag{9}$$

La définition de la norme d'un opérateur linéaire entraîne que le vecteur y_0 d'un espace de dimension finie peut toujours être choisi de façon que $\|R_h^n y_0\| = \|R_h^n\| \cdot \|y_0\|$ pour n donné. Pour un certain y_0 (dépendant de h) on a donc

$$\max_n \|R_h^n y_0\| = \max_n \|R_h^n\|_Y \cdot \|y_0\|. \tag{10}$$

Avec ce choix de y_0, on obtient en vertu de (9) et (10)

$$\|u^{(h)}\|_{U_h} \geqslant M_1 \max_n \|R_h^n\| \cdot \|y_0\| \geqslant M_1 M_2 \max_n \|R_h^n\| \cdot \|f^{(h)}\|_{F_h}. \tag{6'}$$

Il découle de la dernière estimation que si le schéma aux différences (1) est stable, la constante C' entrant dans la définition de la stabilité

$$\|u^{(h)}\|_{U_h} < C' \|f^{(h)}\|_{F_h}$$

vérifie nécessairement l'estimation

$$C' \geqslant M_1 M_2 \max_n \|R_h^n\|. \tag{6''}$$

On en conclut en particulier que *si les normes* $\|u^{(h)}\|_{U_h}$, $\|f^{(h)}\|_{F_h}$ *et* $\|y_n\|$ *concordent de façon qu'on ait les conditions* (6) *et* (7), *alors* (3) *est une condition nécessaire de stabilité*. La condition (3) équivaut à assujettir la solution $\{y_n\}$ de l'équation homogène $y_{n+1} = R_h y_n$ à satisfaire pour tout y_0 à l'inégalité

$$\|y_n\| \leqslant C\|y_0\|, \quad n = 1, 2, \ldots, N. \tag{11}$$

Dans les exemples 1 et 2 du §14, les nombres M_1 et M_2 peuvent être indépendants de h (égaux à 1), fait que le lecteur vérifiera sans peine et qui témoigne du caractère naturel des conditions formulées.

Dans l'exemple 3 du §14, où le schéma aux différences (22) a été rendu canonique à l'aide de l'égalité (24) et où les espaces sont munis des normes (23), la condition $\|u^{(h)}\|_{U_h} \geqslant M_1 \max\limits_n \|y_n\|$ n'est remplie que si $M_1 \leqslant h/2$. Pour un autre choix de la norme $\|u^{(h)}\|_{U_h}$, notamment en posant

$$\|u^{(h)}\|_{U_h} = \max\left[\max_n \|u_n\|, \max_n \left|\frac{u_{n+1}-u_n}{h}\right|\right], \tag{12}$$

on peut faire $M_1 = 1$, $M_2 = 1$ et l'estimation (3) est une condition nécessaire de stabilité. Des normes ainsi modifiées satisfont toujours aux conditions (17) du §14 sous lesquelles l'estimation (3) en est une condition suffisante.

2. Critère spectral de stabilité. On évalue également $\|R_h^n\|$ par les valeurs propres de la matrice R_h, i.e. par les racines λ de l'équation

$$\det\|R_h - \lambda E\| = 0.$$

Si λ est valeur propre, il existe un vecteur propre y tel que $R_h y = \lambda y$. C'est pourquoi

$$R_h^n y = \lambda^n y, \quad \|R_h^n y\| = |\lambda|^n \|y\|, \quad \|R_h^n\| \geqslant |\lambda|^n.$$

Pour que $\|R_h^n\|$ soit borné, il est donc nécessaire qu'il en ait de même des puissances des valeurs propres $|\lambda^n|$, $n = 1, 2, \ldots, N$. Pour cela il aut q ue toutes les valeurs propres se situent dans le cercle

$$|\lambda| \leqslant 1 + ch \tag{13}$$

du plan complexe (c étant indépendant de h). Dans le cas contraire, pour c quelconque et h suffisamment petit,

$$\|R_h^n\| \geqslant |\lambda|^N > (1+ch)^{\frac{1}{h}} = e^{\frac{1}{h}\ln(1+ch)} \geqslant e^{c\left(1-\frac{ch}{2}\right)} > e^{\frac{c}{2}}.$$

Le eritère énoncé d'évaluation de $\|R_h^n\|$ selon la position du spectre (i.e. l'ensemble de valeurs propres) de l'opérateur R_h ne dépend évidemment pas de la façon dont on introduit la norme dans l'espace où opère R_h.

Le critère spectral de stabilité (13) ne dépend pas non plus de la technique de réduction du schéma (1) à la forme (2). Si au lieu de (2), on a $y'_{n+1} = R'_h y'_n + h\varrho'_n$, de sorte que $y' = Ty$, $R'_h = TR_hT^{-1}$ avec T un opérateur linéaire non dégénéré arbitraire, alors les spectres de R_h et R'_h coïncident.

En effet,

$$\det(R'_h-\lambda E) = \det(TR_hT^{-1}-\lambda E) = \det[T(R_h-\lambda E)T^{-1}] =$$
$$= \det T \det(R_h-\lambda E)\det T^{-1} = \det(R_h-\lambda E).$$

Aussi les équations $\det(R_h-\lambda E) = 0$ et $\det(R'_h-\lambda E) = 0$ possèdent-elles les racines λ identiques.

3. Discussion du critère spectral de stabilité. Nous avons montré plus haut qu'en choisissant les normes conformément aux conditions (6) et (7), la position du spectre de l'opérateur R_h dans le cercle

$$|\lambda| \leqslant 1+ch, \tag{13}$$

qui est nécessaire pour que $\|R_h^n\|$ soient bornés, constitue de même une condition nécessaire de stabilité.

Supposons la condition (13) enfreinte brutalement de façon que pour des $h > 0$ suffisamment petits, il existe une valeur propre λ de module notablement supérieur à 1, par exemple

$$|\lambda| > 1+h^{1-\varepsilon},$$

où $\varepsilon > 0$ est indépendant de h. Le schéma aux différences (1) est alors instable pour tout choix raisonnable des normes $\|u^{(h)}\|_{U_h}$ et $\|f^{(h)}\|_{F_h}$ même sans imposer à ce choix les conditions (6) et (7).

Ce que nous venons de dire ne constitue pas un théorème ne serait-ce que parce que nous nous sommes servi du terme « raisonnable » n'ayant pas reçu de définition rigoureuse. Voyons ce que nous avons en vue.

Pour tout choix raisonnable de la norme $\|u^{(h)}\|_{U_h}$ on peut prendre un k_1 positif tel que, quels que soient h suffisamment petits, on ait l'inégalité

$$\|u^{(h)}\|_{U_h} \geqslant h^{k_1} \max_n |u_n|. \tag{14}$$

Dans le cas contraire, l'égalité (4) du §13

$$\lim_{h\to 0} \|[u]_h\|_{U_h} = \|u\|_U$$

ne peut évidemment avoir lieu.

Ensuite, pour tout choix raisonnable de la norme $\|f^{(h)}\|_{F_h}$ on peut prendre un $k_2 > 0$ tel que pour tous les h suffisamment petits on ait l'inégalité

$$\|f^{(h)}\|_{F_h} < h^{-k_2}F, \tag{15}$$

où F désigne le maximum des modules des composantes de $f^{(h)} \in F_h$. Dans le cas contraire, le schéma aux différences (1) ne peut approcher le problème $Lu = f$ pour la solution u : nous avons constaté que les composantes du résidu $\delta f^{(h)}$ apparu par substitution de $[u]_h$ dans le premier membre du schéma approximant (1) tendent vers zéro au plus comme une certaine puissance du pas h.

Ramenons maintenant le schéma (1) à la forme (2) en posant

$$y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}, \quad \left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_Y = \max[|\alpha|, |\beta|]. \tag{16}$$

Nous estimons, pour fixer les idées, que l'équation aux différences proposée relie trois valeurs successives u_{n-1}, u_n, u_{n+1}.

Si le second membre de l'équation aux différences à la base du schéma (1) est posé égal à zéro, alors, pour un certain $r > 0$, on a l'inégalité

$$\max[|u_0|, |u_1|] = \|y_0\|_Y \geqslant h^r F, \tag{17}$$

puisque les relations entre u_1 et u_0, qui figurent dans le schéma aux différences, sont de la forme

$$\left.\begin{aligned} u_0 &= a, \\ u_1 &= b, \end{aligned}\right\} \quad \text{ou} \quad \left.\begin{aligned} u_0 &= a, \\ \frac{u_1-u_0}{h} &= b \end{aligned}\right\}$$

ou d'une forme analogue.

On constate donc qu'on a toujours les inégalités (6) et (7) à condition de poser $M_1(h) = h^{k_1}$, $M_2(h) = h^{r+k_2}$. En effet (voir également (14) et (17)),

$$\|u^{(h)}\|_{U_h} \geqslant h^{k_1} \max |u_n| = h^{k_1} \max_n \|y_n\|,$$

$$\|y_0\| \geqslant h^r F = h^r h^{k_2}(h^{-k_2}F) \geqslant h^{r+k_2} \|f^{(h)}\|_{F_h}.$$

Et l'inégalité (6′) s'écrit donc

$$\|u^{(h)}\|_{U_h} \geqslant h^{r+k_1+k_2} \max_n \|R_h^n\| \cdot \|f^{(h)}\|_{F_h} \geqslant h^{r+k_1+k_2}(1+h^{1-\varepsilon})^{1/h} \|f^{(h)}\|_{F_h}.$$

Ce qui équivaut à l'instabilité parce que, quels que soient r, k_1, k_2 et $\varepsilon > 0$, visiblement

$$h^{r+k_1+k_2}(1+h^{1-\varepsilon})^{1/h} \to \infty \quad \text{pour } h \to 0.$$

Nous avons donc montré que s'il y a parmi les valeurs propres de la matrice R_h une racine satisfaisant à l'inégalité $|\lambda| > 1+h^{1-\varepsilon}$, il y a instabilité pour tout choix raisonnable des normes.

Recourons au critère spectral nécessaire de stabilité (13) pour démontrer l'instabilité du schéma du §9. (Dans ce paragraphe, il nous a été impossible de nous livrer à une étude rigoureuse de l'instabilité ne serait-ce qu'à défaut de définitions exactes.)

Le schéma aux différences dont nous parlons approche le problème

$$\left.\begin{aligned} u'+Au &= 0, \quad 0 \leqslant x \leqslant 1, \\ u(0) &= a \end{aligned}\right\} \tag{18}$$

et est de la forme

$$\left.\begin{aligned} 4\frac{u_{n+1}-u_{n-1}}{2h} - 3\frac{u_{n+1}-u_n}{h} + Au_n &= 0, \quad n = 1, 2, \ldots, N+1, \\ u_0 &= a, \\ u_1 &= (1-Ah)a. \end{aligned}\right\} \tag{19}$$

En posant $y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}$ nous ramenons le schéma (19) à la forme (2), où

$$R_h = \begin{pmatrix} 3+Ah & -2 \\ 1 & 0 \end{pmatrix}, \quad \varrho_n \equiv 0.$$

Les valeurs propres de la matrice R_h sont les racines de l'équation du deuxième degré $\det(R_n - \lambda E) = 0$:

$$\lambda_{1,2} = \frac{3+Ah}{2} \pm \sqrt{\left(\frac{3+Ah}{2}\right)^2 - 2}.$$

La racine $\lambda_1(h)$ tend pour $h \to 0$ vers 2 si bien que, lorsque h est petit,

$$|\lambda_1| > \frac{3}{2} > 1.$$

Il n'y aura donc stabilité pour aucun choix raisonnable des normes.

En particulier, si les normes sont introduites par les égalités

$$\|u^{(h)}\|_{U_h} = \max_n |\varphi_n|,$$

$$\|f^{(h)}\|_{F_h} = \left\| \begin{matrix} \varphi_n \\ \alpha \\ \beta \end{matrix} \right\|_{F_h} = \max\left[|\alpha|, |\beta|, \max_n |\varphi_n|\right],$$

$$\|y_n\|_Y = \left\| \begin{bmatrix} y_n^{(1)} \\ y_n^{(2)} \end{bmatrix} \right\|_Y = \max\left[|y_n^{(1)}|, |y_n^{(2)}|\right],$$

on vérifie les deux conditions (6), (7) sous lesquelles l'inégalité (3) est une condition nécessaire de stabilité. Or, $\|R_h^n\| > (^3/_2)^n \to \infty$ si $n = 1/h$, $h \to 0$, et la stabilité n'a pas lieu.

On a vu qu'une violation grossière du critère spectral nécessaire de stabilité (13) :

$$|\lambda| \leqslant 1+ch,$$

par exemple l'existence d'une valeur propre λ^* de l'opérateur R_h évaluée par

$$|\lambda^*| > 1+h^{1-\varepsilon},$$

atteste une instabilité qui ne devient stabilité pour aucun choix des normes.

Soulignons néanmoins que le fait que le spectre de R_h se trouve dans le cercle $|\lambda| < 1+ch$ ne suffit pas pour garantir la stabilité. Celle-ci peut en l'occurrence dépendre d'un choix heureux des normes. Servons-nous, pour le montrer, du schéma aux différences que nous avons déjà examiné au §14 (d'un point de vue différent).

Ecrivons comme suit le schéma aux différences pour le problème $u'' = \varphi(x)$, $u(0) = a$, $u'(0) = b$:

$$\left.\begin{aligned} \frac{u_{n+1}-2u_n+u_{n-1}}{h^2} &= \varphi_n, \quad n = 1, 2, \ldots, N-1, \\ u_0 &= a, \\ \frac{u_1-u_0}{h} &= b. \end{aligned}\right\}$$

En posant $y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}$ nous le mettons sous la forme (2) avec

$$R_h = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}, \quad \varrho_n = \begin{bmatrix} h\varphi_n \\ 0 \end{bmatrix}, \quad y_0 = \begin{bmatrix} a+hb \\ a \end{bmatrix}.$$

La matrice R_h a ses deux valeurs propres égales à l'unité. Si $\varphi_n \equiv 0$, la solution $\{u_n\}$ du problème est de la forme

$$u_n = u_0 + (u_1 - u_0)n, \quad n = 0, 1, 2, \ldots, N.$$

Utilisons deux jeux de normes

1)
$$\|y_n\|_Y = \max(|y_n^{(1)}|, |y_n^{(2)}|),$$
$$\|u^{(h)}\|_{U_h} = \max_n \|y_n\|_Y,$$
$$\|f^{(h)}\|_{F_h} = \left\| \begin{matrix} \varphi_n \\ a \\ b \end{matrix} \right\|_{F_h} = \max\left[\|y_0\|_Y, \max_m |\varphi_m|\right];$$

2) $$\|y_n\|_{Y_h} = \max\left(|y_n^{(1)}|, \left|\frac{y_n^{(2)} - y_n^{(1)}}{h}\right|\right), \quad \|u^{(h)}\|_{U_h} = \max_n \|y_n\|_{Y_h},$$
$$\|f^{(h)}\|_{F_h} = \max\left[\|y_0\|_{Y_h}, \max_m |\varphi_m|\right].$$

Le lecteur s'assurera facilement que dans les deux cas les conditions (6), (7) et les conditions (28) du §14 sous lesquelles la stabilité équivaut à l'estimation

$$\|R_h^n\| \leqslant C, \quad n = 1, 2, \ldots, N-1,$$

se trouvent satisfaites.

Avec les normes définies par les formules 1) cette estimation n'a pas lieu. En posant par exemple $y_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\|y_0\| = 1$, on obtient

$$y_n = \begin{bmatrix} n+1 \\ n \end{bmatrix}, \quad \|y_n\| = n+1, \quad \|R_h^n\| \geqslant n+1 \to \infty$$

pour $n = 1/h$, $h \to 0$.

Les normes étant choisies selon les formules 2), il y a stabilité : on a pour $y_0 = \begin{bmatrix} u_1 \\ u_0 \end{bmatrix}$ arbitraire

$$\|y_n\|_{Y_h} = \|R_h^n y_0\| = \left\| \begin{matrix} u_0 + (u_1 - u_0)(n-1) \\ u_0 + (u_1 - u_0)n \end{matrix} \right\|_{Y_h} =$$
$$= \max\left[|u_0 + (u_1 - u_0)(n+1)| \cdot \left|\frac{u_1 - u_0}{h}\right|\right].$$

Mais $n+1 \leqslant 1/h$, donc

$$\|y_n\|_{Y_h} = \|R_h^n y_0\|_{Y_h} \leqslant |u_0| + \left|\frac{u_1 - u_0}{h}\right| \leqslant 2\|y_0\|_{Y_h}$$

et

$$\|R_h^n\| \leqslant 2, \quad n = 1, 2, \ldots, N-1.$$

Dans la pratique, on se borne souvent à tester le critère spectral nécessaire de stabilité. Si ce critère est vérifié, on teste le schéma par le calcul expérimental suivant ce schéma même sans construire explicitement les normes. Cette vérification fait la matière du §18.

Exercices

1. Supposons l'équation aux différences du second ordre $au_{n-1}+bu_n+cu_{n+1} = \varphi_n$ réduite à la forme $y_{n+1} = R_h y_n + h\varphi_n$ par la substitution $y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}$. Montrer que les racines de l'équation caractéristique $a+b\lambda+c\lambda^2 = 0$ coïncident avec les valeurs propres de la matrice R_h.

2. Ecrire l'équation aux différences du second ordre $au_{n-1}+bu_n+cu_{n+1} = \varphi_n$ sous la forme $y_{n+1} = R_h y_n + h\varrho_n$ moyennant la substitution $y_n = \begin{bmatrix} u_{n+2} \\ u_{n+1} \\ u_n \end{bmatrix}$. Une telle réduction est-elle unique ? Montrer que les valeurs propres de la matrice R_h consistent en les racines de l'équation caractéristique $a+b\lambda+c\lambda^2 = 0$ plus le nombre $\lambda = 0$, de sorte que le critère spectral de stabilité $|\lambda| \leqslant 1+ch$ a lieu indifféremment pour $y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}$ et $y_n = \begin{bmatrix} u_{n+2} \\ u_{n+1} \\ u_n \end{bmatrix}$.

3. Les vecteurs propres $v^{(1)}$ et $v^{(2)}$ d'une matrice R_h d'ordre 2 associés aux valeurs propres respectives λ_1 et λ_2

$$R_h v^{(1)} = \lambda_1 v^{(1)}, \qquad R_h v^{(2)} = \lambda_2 v^{(2)}$$

tendant pour $h \to 0$ vers des positions limites non colinéaires différentes, les conditions $|\lambda_1| < 1+ch$, $|\lambda_2| < 1+ch$ sont non seulement nécessaires mais aussi suffisantes pour avoir une estimation de la forme $\| R_h^n \| < C$, $n = 1, 2, \ldots, N$, si $\left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_T = \max [|\alpha|, |\beta|]$. Démontrer.

§16. Erreurs d'arrondi

1. Erreurs sur les coefficients. Si le schéma aux différences

$$L_h u^{(h)} = f^{(h)} \tag{1}$$

approche le problème $Lu = f$ pour la solution u et est stable, il y a convergence. Jamais la réalisation du schéma conçu n'est cependant exacte vu les erreurs d'arrondi à la donnée de ses coefficients et seconds membres.

On demande par exemple de résoudre le problème

$$\left.\begin{aligned} u'+Au &= \cos x, \quad 0 \leqslant x \leqslant 1, \\ u(0) &= a \end{aligned}\right\}$$

en recourant au schéma aux différences

$$\left.\begin{aligned} \frac{u_{n+1}-u_n}{h}+Au_n &= \cos x_n, \quad n = 0, 1, \ldots, N-1, \\ u_0 &= a. \end{aligned}\right\} \tag{2}$$

Les valeurs de $\cos x_n$, a, A et le coefficient $1/h$ sont donnés avec certain arrondi. Dans le cas général, on a affaire, au lieu de (1), avec le schéma

$$L_h v^{(h)} + (\Delta^{(h)} L_h) v^{(h)} = f^{(h)} + \Delta^{(h)} f^{(h)}, \tag{3}$$

$\Delta^{(h)} L_h$ et $\Delta^{(h)} f^{(h)}$ étant des erreurs d'arrondi à la donnée de l'opérateur L_h et du second membre $f^{(h)}$. L'opérateur $\Delta^{(h)} L_h$ du schéma (2) a la forme

$$(\Delta^{(h)} L_k) v^{(h)} = \begin{cases} \Delta^{(h)}\left(\frac{1}{h}\right)(v_{n+1} - v_n) + (\Delta^{(h)} A) v_n, & n = 0, 1, \ldots, N-1, \\ 0 \cdot v_0. \end{cases}$$

L'erreur $\Delta^{(h)} f^{(h)}$ est définie par la formule

$$\Delta^{(h)} f^{(h)} = \begin{cases} \Delta^{(h)} \cos x_n, & n = 0, 1, \ldots, N-1, \\ \Delta^{(h)} a. \end{cases}$$

Ici $\Delta^{(h)} M$ est l'erreur commise en se donnant M.

Afin d'éviter des difficultés purement techniques, bornons-nous au cas des opérateurs L_h, $\Delta^{(h)} L_h$ linéaires et de l'espace U_h de dimension finie comme pour le schéma (2). Etablissons sous ces hypothèses les erreurs d'arrondi permises et la loi de croissance de la précision sur la donnée du schéma avec le resserrement du réseau, i.e. pour h tendant vers zéro.

THÉORÈME. *Si le schéma aux différences stable* (1) *approche le problème* $Lu = f$ *pour la solution* u *à un ordre* h^k :

$$\| L_h [u]_h - f^{(h)} \|_{F_h} \leqslant c h^k,$$

alors dans les conditions

$$\left.\begin{aligned} \| \Delta^{(h)} L_h v^{(h)} \|_{F_h} &\leqslant c_1 h^k \| v^{(h)} \|_{U_h}, \\ \| \Delta^{(h)} f^{(h)} \|_{F_h} &\leqslant c_2 h^k \end{aligned}\right\} \tag{4}$$

le schéma aux différences (3) *approche lui aussi le problème* $Lu = f$ *à l'ordre* h^k *et est stable.*

Ainsi, dans les conditions (4), l'ordre de précision du schéma (3), support effectif du calcul, est h^k et il coïncide avec celui du schéma conçu (1).

Sous l'hypothèse de la norme $\|\cdot\|_{U_h}$ choisie en conformité avec la condition (4) du § 13, i.e. de façon que

$$\lim_{h \to 0} \| [u]_h \|_{U_h} = \| u \|_U,$$

la quantité $\| [u]_h \|_{U_h}$ reste bornée pour $h \to 0$, $\| [u]_h \|_{U_h} \leqslant P < \infty$. Notons

$$\tilde{L}_h u^{(h)} \equiv L_h u^{(h)} + (\Delta^{(h)} L_h) u^{(h)},$$
$$\tilde{f}^{(h)} \equiv f^{(h)} + \Delta^{(h)} f^{(h)}$$

et assurons-nous que le schéma $\tilde{L}_h u^{(h)} = \tilde{f}^{(h)}$ donne une approximation

de l'ordre de h^k. En effet, on a

$$\|\tilde{L}_h[u]_h-\tilde{f}^{(h)}\|_{F_h} = \|L_h[u]_h-f^{(h)}+(\Delta^{(h)}L_h[u]_h-\Delta^{(h)}f^{(h)})\|_{F_h} \leqslant$$
$$\leqslant \|L_h[u]_h-f^{(h)}\|_{F_h}+\|\Delta^{(h)}L_h[u]_h\|_{F_h}+\|\Delta^{(h)}f^{(h)}\|_{F_h} \leqslant ch^k+c_1Ph^k+c_2h^k \leqslant \tilde{c}h^k.$$

La démonstration du théorème fera appel au lemme connu suivant.

LEMME. *Soient A et B deux opérateurs linéaires d'un espace vectoriel normé de dimension finie X dans un espace vectoriel normé G.*

Supposons ensuite que pour un $g \in G$ quelconque il existe une solution $x \in X$ de l'équation

$$Ax = g,$$

que

$$\|x\|_X \leqslant c\,\|g\|_G \tag{5}$$

et que tout $x \in X$ vérifie l'inégalité

$$\|Bx\|_G \leqslant \frac{q}{c}\,\|x\|_X, \tag{6}$$

où $c > 0$, $0 < q < 1$, c et q étant des nombres. L'équation

$$(A+B)\tilde{x} = g$$

admet alors pour tout $g \in G$ une solution unique et

$$\|\tilde{x}\|_X \leqslant \frac{c}{1-q}\,\|g\|_G. \tag{7}$$

DÉMONSTRATION. Notons que X et G sont de même dimension, sinon il y aurait des $g \in G$ tels que le problème $Ax = g$ soit impossible.

Si x_0 est une solution de

$$(A+B)x_0 = g,$$

alors

$$Ax_0 = g-Bx_0,$$
$$x_0 = A^{-1}g-A^{-1}Bx_0,$$

où $A^{-1}g$ et $A^{-1}Bx_0$ sont solutions des équations $Ax = g$ et $Ax = Bx_0$,

$$\|x_0\|_X \leqslant \|A^{-1}g\|_X+\|A^{-1}(Bx_0)\|_X \leqslant c\,\|g\|_G+c\,\|Bx_0\| \leqslant$$
$$\leqslant c\,\|g\|_G+c\,\frac{q}{c}\,\|x_0\|_X.$$

D'où

$$\|x_0\|_X \leqslant \frac{c}{1-q}\,\|g\|_G.$$

La dernière inégalité entraîne que pour $g = 0$ l'équation $(A+B)x = g$ ne possède comme solution que la solution triviale $x_0 = 0$, d'où l'existence d'une solution unique pour $g \in G$ arbitraire et l'estimation (7).

DÉMONSTRATION DU THÉORÈME. Utilisons le lemme et prenons pour opérateurs A et B respectivement L_h et $\Delta^{(h)}L_h$. L'existence d'une solution du problème $Ax = g$ et l'estimation (5) équivalent à supposer le schéma (1) stable. L'estimation (6) est valide en vertu de (4) pour tout q positif à condition d'avoir un h suffisamment petit.

La résolubilité de l'équation $(A+B)x = g$ pout tout $g \in G$ et l'estimation (7) équivalent exactement à la stabilité du schéma aux différences (3).

Notons que les contraintes (4) imposées aux erreurs d'arrondi à la donnée d'un schéma aux différences stable sont bien raisonnables : si, en diminuant h, nous voulons une réponse à h^k près, i.e. possédant des chiffres décimaux en nombre de l'ordre de $\ln(1/h)$, il faut améliorer continûment la précision sur les coefficients du schéma en augmentant le nombre de chiffres des valeurs données à une vitesse de l'ordre de $\ln(1/h)$. Cette augmentation est en général parfaitement possible, $\ln(1/h)$ étant une fonction à croissance lente. Par contre, on n'arrive pas à élever la précision en diminuant les pas sans augmenter le nombre de chiffres décimaux avec lesquels sont donnés les coefficients et les seconds membres.

2. Erreurs de calcul. Etant donné un schéma aux différences, il faut calculer sa solution $u^{(h)}$. Supposons que nous savons résoudre exactement les équations aux différences. Si le schéma aux différences mis en œuvre approche l'équation différentielle et jouit de la stabilité, pour un pas suffisamment petit, la solution $u^{(h)}$ diffère peu de la solution exacte cherchée $[u]_h$.

Cela étant, l'ordre d'opérations (l'algorithme) utilisé pour calculer $u^{(h)}$ est sans importance vu que le résultat en est indépendant.

Or, nous commettons en réalité à chaque pas de l'algorithme choisi des erreurs d'arrondi qui influent sur les résultats obtenus aux pas suivants. Le pas h étant fixe et l'espace U_h de dimension finie, l'algorithme consiste en une suite finie d'opérations arithmétiques (sommation, division, soustraction, produit), et le résultat de chaque opération dépend de façon continue des opérandes. Aussi, en travaillant avec un nombre « suffisamment grand » de décimales, nous pouvons calculer $u^{(h)}$ avec un nombre de décimales donné d'avance. Le nombre de décimales « de réserve » avec lesquelles on travaille en vue d'obtenir un $u^{(h)}$ possédant des chiffres en nombre donné dépend et de l'algorithme choisi et de h. Ainsi, nous avons montré plus haut (§7) qu'en résolvant par balayage un problème aux limites bien conditionné le nombre de décimales « de réserve » ne croît nullement avec $h \to 0$. Bien que paraissant raisonnables, certains algorithmes de résolution des problèmes stables exigent parfois que le nombre de décimales « de réserve » croisse vite proportionnellement à $1/h$. Un exemple en est donné au n° 2 du §5. Ce nombre doit en général augmenter avec la diminution de h. Les algorithmes dans lesquels cette croissance est trop rapide sont considérés comme instables et numériquement impraticables. L'étude de la stabilité des algorithmes est un problème compliqué. Citons comme exemple d'une telle étude la justification du balayage (§ 7). Dans des cas élémentaires on arrive cependant à établir le nombre de décimales « de réserve » nécessaire au calcul connaissant la propriété de stabilité du

schéma et le théorème sur la possibilité de définir ce schéma de façon approchée (n° 1).

Soit par exemple le schéma

$$\frac{u^{(h)}(x+h)-u^{(h)}(x)}{h}+Au^{(h)}(x) = f^{(h)}(x).$$

En calculant $u^{(h)}(x+h)$ par la formule récurrentielle

$$u^{(h)}(x+h) = u^{(h)}(x)(1-Ah)+hf^{(h)}(x)$$

et en travaillant avec un nombre fini de décimales, nous pouvons commettre une erreur δ sur $u^{(h)}(x+h)$. Il y a intérêt à estimer qu'il s'agit d'une erreur non pas sur cette dernière valeur, mais sur le second membre $f^{(h)}(x)$, ce qui revient à dire que $u^{(h)}(x+h)$ est calculé exactement et qu'au lieu de $f^{(h)}(x)$ nous avons utilisé la quantité $f^{(h)}(x)+\delta/h$. Du moment que nous commettons de telles erreurs en tous les points x, la quantité δ est supposée dépendre de x: $\delta = \delta(x)$. Dans cet exemple, l'erreur qu'on commet en arrondissant lors du calcul peut être prise pour erreur $\delta(x)/h$ à la donnée du second membre. Le schéma considéré permet une approximation au premier ordre et est stable. Pour ne pas altérer la convergence en h, on doit donc augmenter sans cesse la précision des calculs, à savoir de façon que

$$\Delta^{(h)}f^{(h)} = \frac{\delta(x)}{h}$$

soit de l'ordre de h.

Pour cela $\delta(x)$ doit être en h^2. On y arrive en calculant $u^{(h)}$ avec un nombre de décimales de réserve croissant comme $\ln(1/h)$ pour $h \to 0$.

Nous avons montré dans cet exemple que dans des cas simples on peut regarder les erreurs d'arrondi commises en calculant la solution $u^{(h)}$ à un facteur de la forme h^m près comme les erreurs à la donnée des seconds membres $f^{(h)}$. Le théorème démontré plus haut entraîne que s'agissant de schémas stables ces erreurs n'entravent pas la convergence sans perte de précision si le nombre de décimales sur lesquelles on opère croît lentement comme $c \ln(1/h)$, c étant une constante.

§17. Caractérisation quantitative de la stabilité

Examinons d'abord le schéma aux différences bien connu

$$\left.\begin{aligned} \frac{u_{n+1}-u_n}{h}+Au_n &= 0, \\ u_0 &= 1 \end{aligned}\right\} \qquad (1)$$

pour le problème aux limites différentiel

$$u'+Au = 0, \quad u(0) = 1.$$

Sa solution est de la forme

$$u_n = e^{-Ax_n} + h\frac{A^2x_n}{2}e^{-Ax_n} + O(h^2)$$

(voir (3′) du §8; on pose $b = 1$). L'expression (6) du §8

$$\delta(x_n) = h\frac{A^2x_n}{2}e^{-Ax_n} + O(h^2)$$

représente un reste, i.e. l'erreur faite en remplaçant la valeur e^{-Ax_n} de la solution exacte de l'équation différentielle par la solution $u_n^{(h)}$ du problème aux différences. Le reste tend vers zéro comme la première puissance de h ; ce schéma est exact à l'ordre 1. Le choix du pas h dépend de la précision désirée. Il est clair que le module du rapport de l'erreur à la solution exacte $|\delta(x_n)/u(x_n)|$ doit être en tout cas inférieur à 1 pour qu'on puisse parler d'une précision quelconque sur la solution approchée.

Voyons quels h satisfont à cette condition. Négligeons le terme $O(h^2)$ dans l'expression de $\delta(x_n)$ et étudions le rapport de l'erreur $\delta(x_n)$ au point x_n à la solution exacte :

$$\frac{\delta(x_n)}{u(x_n)} \approx \frac{h\dfrac{A^2x_n}{2}e^{-Ax_n}}{e^{-Ax_n}} = h\frac{A^2x_n}{2}.$$

Faisons $A = 20$ et $x_n = 1$. Nous obtenons alors, à partir de la condition $|\delta(1)/u(1)| < 1$:

$$h < 0{,}2\cdot 10^{-3}.$$

Etablissons les pas avec lesquels on intègre le même problème $u' + Au = 0$ moyennant le schéma

$$\left.\begin{aligned} \frac{u_{n+1}-u_{n-1}}{2h} + Au_n &= 0, \\ u_0 &= 1, \\ u_1 &= 1 - Ah \end{aligned}\right\} \tag{2}$$

exact à l'ordre deux si $A = 20$ et si on veut remplir la condition

$$\left|\frac{\delta(1)}{u(1)}\right| < 1. \tag{3}$$

La solution de ce problème est de la forme (voir (12) §8 pour $b = 1$) :

$$u_n = e^{-Ax_n} + h^2\left[\frac{2Ax_n-3}{12}A^2e^{-Ax_n} + (-1)^n\frac{A^2}{4}e^{Ax_n}\right] + O(h^3),$$

et l'erreur s'écrit donc

$$\delta(x_n) = h^2\left[\frac{2Ax_n-3}{12}A^2e^{-Ax_n} + (-1)^n\frac{A^2}{4}e^{Ax_n}\right] + O(h^3).$$

Négligeons le terme $O(h^3)$, écrivons le rapport de l'erreur $\delta(x_n)$ à la solution exacte $u(x_n) = e^{-Ax_n}$ et définissons le pas h par la condition (3). Ce dernier

s'avérera si petit qu'en assimilant conventionnellement à 1 seconde le temps machine de calcul selon le schéma (1), le calcul par le schéma (2) demandera près de quatre jours.

Le fait est que la validité d'un schéma face à un certain problème se mesure non seulement à la puissance de h figurant dans la formule donnant l'erreur, mais aussi au coefficient de cette puissance.

Voyons maintenant comment la discussion de la stabilité d'un schéma aux différences $L_h u^{(h)} = f^{(h)}$ permet de juger de sa validité. Pour abréger l'écriture l'opérateur L_h sera suupposé linéaire. Rappelons (voir § 12) qu'un schéma aux différences est dit stable si, quel que soit $f^{(h)} \in F_h$, il admet une solution unique, la solution $u^{(h)} \in U_h$ étant évaluée par

$$\|u^{(h)}\|_{U_h} \leqslant C\,\|f^{(h)}\|_{F_h}.$$

En démontrant un théorème aux termes duquel l'approximation et la stabilité entraînent la convergence (§ 12), nous avons obtenu pour l'erreur $z^{(h)} = [u]_h - u^{(h)}$ l'inégalité suivante :

$$\|z^{(h)}\|_{U_h} \leqslant CC_1h^k, \tag{4}$$

C_1h^k évaluant l'erreur d'approximation

$$\|L_h[u]_h - f^{(h)}\|_{F_h} \leqslant C_1h^k.$$

Soit C_1h^k petit. L'estimation de $\|z^{(h)}\|_{U_h}$ montre que la petitesse de $\|[u]_h - u^{(h)}\|_{U_h}$ exige de plus que le coefficient C caractérisant la stabilité ne soit pas trop grand.

Si nous voulons donc vérifier si tel ou tel schéma aux différences peut être utilisé pour résoudre le problème proposé, le fait de le savoir stable ne suffit pas. Il faut de plus connaître approximativement la grandeur du coefficient C, et on y arrive par les techniques indiquées aux §§ 14, 15, par des calculs expérimentaux ou par un procédé indirect.

Calculons par exemple C pour les schémas aux différences (1) et (2) relatifs au problème $u' + Au = \varphi(x)$, $u(0) = a$.

Considérons d'abord le schéma

$$\frac{u_{n+1} - u_n}{h} + Au_n = \varphi_n, \quad n = 0, 1, \ldots, N-1,$$
$$u_0 = a$$

pour les normes

$$\|u^{(h)}\|_{U_h} = \max_n |u_n|, \quad \|f^{(h)}\|_{F_h} = \max\left[|a|, \max_n |\varphi_n|\right].$$

Réduisons-le à la forme

$$\left.\begin{aligned} &y_{n+1} = R_h y_n + h\varrho_n, \\ &y_0 \text{ donné} \end{aligned}\right\}$$

en posant $y_n = u_n$, $R_h = (1 - Ah)$, $\varrho_n = \varphi_n$. Posons $\|y_n\| = |y_n|$. Les

conditions (17) du § 14 se trouvent alors remplies :

$$\left.\begin{aligned} \|u^{(h)}\|_{U_h} &\leqslant C_2 \max_n \|y_n\|, \\ \|\varrho_n\| &\leqslant C_2 \|f^{(h)}\|_{F_h}, \\ \|y_0\| &\leqslant C_2 \|f^{(h)}\|_{F_h}, \end{aligned}\right\} \tag{5}$$

et on peut faire $C_2 = 1$.

Il est évident que $\|R_h^n\| = (1-Ah)^n$. On peut donc poser $C = 2\max[1, (1-Ah)^n]$. D'où

$$C = \begin{cases} 2 & \text{si } A > 0, \\ 2(1-Ah)^N & \text{si } A \leqslant 0. \end{cases}$$

Montrons que le nombre C ne peut être sensiblement inférieur. Les normes sont choisies de façon qu'on ait également les conditions (6) et (7) du § 15 :

$$\|u^{(h)}\|_{U_h} \geqslant M_1 \max_n \|y_n\|, \tag{6}$$

et, pour $\varphi_n = 0$ ($\varrho_n = 0$),

$$\|y_0\| \geqslant M_2 \|f^{(h)}\|_{F_h}; \tag{7}$$

en outre on peut poser $M_1 = M_2 = 1$. C'est pourquoi la constante C doit vérifier (on l'a établi au § 15) l'estimation $C \geqslant M_1 M_2 \max \|R_h^n\|$:

$$C \geqslant \begin{cases} 1 & \text{si } A > 0, \\ (1-Ah)^N & \text{si } A \leqslant 0. \end{cases}$$

Evaluons la constante C figurant dans la définition de la stabilité $\|u^{(h)}\|_{U_h} \leqslant C \|f^{(h)}\|_{F_h}$ pour le schéma (2). Ecrivons-le comme

$$\left.\begin{aligned} &y_{n+1} = R_h y_n + h\varrho_n, \quad n = 0, 1, \ldots, \\ &y_0 \text{ donné} \end{aligned}\right\}$$

en posant

$$y_n = \begin{bmatrix} u_{n+1} \\ u_n \end{bmatrix}, \qquad \varrho_n = \begin{bmatrix} 2\varphi_n \\ 0 \end{bmatrix},$$

$$R_h = \begin{pmatrix} -2Ah & 1 \\ 1 & 0 \end{pmatrix}, \quad y_0 = \begin{bmatrix} (1-Ah) \\ 1 \end{bmatrix}.$$

Introduisons les normes

$$\|u^{(h)}\|_{U_h} = \max_n |u_n|,$$

$$\|f^{(h)}\| = \left\| \begin{bmatrix} \varphi_n \\ \alpha \\ \beta \end{bmatrix} \right\| = \max\left[|\alpha|, |\beta|, \max_n |\varphi_n|\right],$$

$$\|y_n\| = \left\| \begin{bmatrix} y_n^{(1)} \\ y_n^{(2)} \end{bmatrix} \right\| = \max\left[|y_n^{(1)}|, |y_n^{(2)}|\right].$$

On a alors les conditions (5) à (7) et $C_2 = M_1 = M_2 = 1$. En vertu du n° 3 du § 14 on prend donc pour C le nombre $C = 2C_2^2 \max_n \|R_h^n\| = 2 \max_n \|R_h^n\|$, mais aux termes de (6″) du § 15 ce nombre peut être diminué de moitié au plus : on doit nécessairement avoir

$$C \geqslant M_1 M_2 \max_n \|R_h^n\| = \max_n \|R_h^n\|.$$

La quantité $\max \|R_h^n\|$ a été majorée au § 14 :

$$\|R_h^n\| \leqslant \left\| \begin{pmatrix} -2Ah & 1 \\ 1 & 0 \end{pmatrix} \right\|^N \leqslant (1+2|A|h)^N.$$

On peut donc poser

$$C = 2e^{2|A|} \geqslant 2(1+2|A|h)^{1/h}.$$

La minoration de $\max \|R_h^n\|$ est obtenue à partir de la condition $\|R_h^n\| \geqslant |\lambda|^n$, λ étant la plus grande en module des deux valeurs propres de la matrice R_h. La résolution de l'équation $\det(R_h - \lambda E) = 0$ fournit les valeurs propres

$$\lambda_1 = 1 - Ah + \frac{A^2h^2}{2} + o(h^2) = 1 - Ah + O(h^2),$$

$$\lambda_2 = -1 - Ah - \frac{A^2h^2}{2} + o(h^2) = -1 - Ah + O(h^2),$$

de sorte que

$$\max(|\lambda_1|, |\lambda_2|) = 1 + |A|h + O(h^2),$$

$$\max_n \|R_h^n\| = (1+|A|h)^{1/h} + O(h).$$

Il est donc notoirement impossible de remplacer la constante $C = 2e^{2|A|}$ obtenue plus haut par un nombre inférieur à $(1+|A|h)^{1/h} \approx e^{|A|}$, i.e. de la diminuer sensiblement.

Lorsque $A = 20$, on a $C = 2$ pour le premier schéma et forcément $C \geqslant e^{20} \geqslant 10^8$ pour le second.

Lorsque $A \approx 1$ ou $A < 0$, les propriétés de stabilité des deux schémas ne diffèrent que légèrement : la constante C est sensiblement la même dans les deux cas.

On comprend facilement pourquoi, lorsque $A \gg 1$, la constante C est notablement supérieure à 1 pour le second schéma et égale à 2 pour le premier.

La solution générale de l'équation homogène $u_{n+1} - (1-Ah)u_n = 0$ correspondant au schéma (1) est

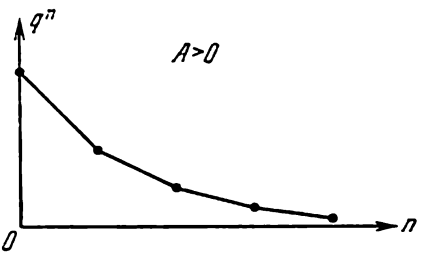


Fig. 4

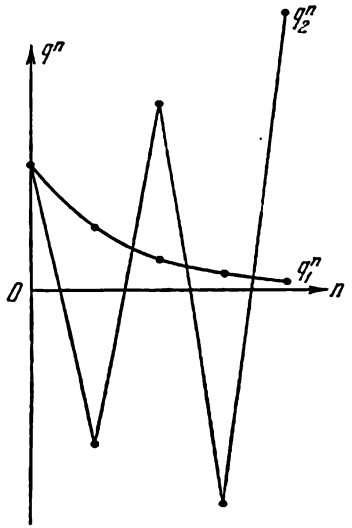


Fig. 5

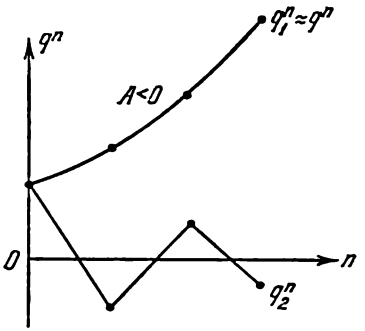


Fig. 6

$\bar{u}_n = \alpha q^n$, q étant racine de l'équation caractéristique $q-(1-Ah)=0$, $q = 1-Ah$ (fig. 4). La solution générale de l'équation homogène

$$u_{n+1}+2Ahu_n-u_{n-1} = 0,$$

correspondant au schéma (2) est

$$\bar{u}_n = \alpha q_1^n+\beta q_2^n,$$

où q_1 et q_2 sont deux racines de l'équation caractéristique

$$q^2+2Ahq-1 = \det(R_h-qE) = 0,$$

$$q_1 = 1-Ah+\frac{A^2h^2}{2}+o(h^2),$$

$$q_2 = -1-Ah-\frac{A^2h^2}{2}+o(h^2).$$

La racine q_1 « ressemble » à la racine $q = 1-Ah$, et il lui correspond la solution q_1^n ressemblant à la solution q^n de la première équation. Mais la racine « parasite » $q_2 = -1-Ah+O(h^2)$ donne la solution « parasite » à croissance rapide q_2^n (fig. 5) qui détermine de grandes valeurs de C.

Pour A négatif, on a $q > 1$, $q_1 > 1$, $|q_2| < 1$. Les solutions q^n et q_1^n correspondant aux racines q et q_1 accusent une croissance rapide quasi identique, et la solution parasite q_2^n s'amortit sans influencer la stabilité du second schéma (fig. 6).

Notons que pour tout schéma aux différences approchant le problème $u'+Au = 0$, $u(0) = a$, C est inévitablement important lorsque $A \ll 0$. En effet, pour h petit, la solution d'un problème aux différences stable ressemble à celle du problème différentiel, qui en est limite quand $h \to 0$. Mais la solution $u = u_0e^{-Ax}$ du problème différentiel est telle que $\max|u(x)| = |u_0|e^{-Ax}$, i. e. $\max|u(x)|$ est supérieur un grand nombre e^{-A} de fois au module $|u_0|$ de la valeur initiale u_0.

Une autre remarque à faire : les valeurs élevées du coefficient C impliquent non seulement un calcul à pas faible, mais aussi un grand nombre de décimales sur lesquelles on opère.

En effet, nous avons montré au § 16 que les erreurs d'arrondi peuvent être incluses dans des erreurs qu'on fait en se donnant les seconds membres et qui sont évaluées par C_1h^k. Le coefficient C_1 croît avec ces erreurs, ce qui, C étant grand, peut, en vertu de (4), s'avérer catastrophique pour le résultat.

Nous terminons ce paragraphe en mettant en garde une fois de plus contre l'idée fausse sur les schémas exacts à l'ordre 2 qui peut être suggérée au lecteur par l'exemple considéré. Car il n'est nullement dans nos intentions de jeter le discrédit sur tous ces schémas en décrivant les défauts de l'un d'eux. Le lecteur ferait bien d'étudier lui-même un schéma précis à l'ordre 2 de la forme

$$\left.\begin{aligned} \frac{u_{n+1}-u_n}{h}+A\frac{u_{n+1}+u_n}{2} &= 0, \\ u_0 &= 1. \end{aligned}\right\}$$

En procédant de façon que pour $A = 1$ l'erreur $\delta(1)$ soit inférieure à $u(1) = e^{-A}$, il se convaincra que ce schéma impose au pas h une contrainte plus lâche que le schéma exact à l'ordre un (1).

Nous conseillons également au lecteur de tâcher de déterminer le pas avec lequel il fait intégrer le problème $u'+u = 0$, $u(0) = 1$, pour obtenir $u(1)$ faux de 10^{-5} au plus. En faisant ces essais pour les schémas aux différences (1) et (2), il verra que le schéma (1) (précision d'ordre 1) exige un pas sensiblement plus petit que (2) (schéma exact à l'ordre 2).

La qualité d'un schéma est donc plus que sa propriété intrinsèque, elle est également fonction du problème qu'il sert à résoudre.

§18. Un procédé d'étude de la stabilité des problèmes non linéaires

En étudiant la stabilité aux §§14 et 15 nous nous sommes servi de techniques propres à des équations aux différences à coefficients constants. On pourrait donc penser qu'on ne saurait utiliser la matière de ces paragraphes pour analyser les schémas d'intégration même des équations $\frac{du}{dx} = G(x, u)$, G étant une fonction assez générale. Or, il n'en est rien.

Supposons que la courbe intégrale cherchée de l'équation

$$\frac{du}{dx} = G(x, u) \tag{1}$$

passe par un point de coordonnées $x = x_0$, $u = u_0$. Au voisinage de ce point

$$G(x, u) \approx \frac{\partial G(x, u)}{\partial u}(u-u_0)+\frac{\partial G(x, u)}{\partial x}(x-x_0)+G(x_0, u_0), \tag{2}$$

et l'équation (1) peut être remplacée avec une certaine précision par une autre plus simple

$$\frac{du}{dx}-Au = \varphi(x), \tag{3}$$

où

$$A = \left.\frac{\partial G(x, u)}{\partial u}\right|_{\substack{x=x_0\\ u=u_0}},$$

$$\varphi(x) = G(x_0, u_0)+\left.\frac{\partial G(x, u)}{\partial x}\right|_{\substack{x=x_0\\ u=u_0}}(x-x_0)-u_0\left.\frac{\partial G(x, u)}{\partial u}\right|_{\substack{x=x_0\\ u=u_0}}.$$

Il est naturel que les schémas que nous nous proposons de mettre en œuvre pour trouver la solution nécessaire doivent bien intégrer l'équation (3) approchant (1) au voisinage d'un point que traverse la courbe intégrale. La valeur du coefficient A obtenu en linéarisant comme décrit plus haut l'équation de départ diffère certes d'un point de la courbe à l'autre. Ainsi, en sélectionnant un schéma aux différences, on le testera non pas sur une seule valeur de A de l'équation (3), mais sur tout un ensemble de valeurs décrivant suffisamment bien les limites de variation de $\partial G/\partial u$ le long de la courbe intégrale. Dans la majorité écrasante des cas réels cette étude suffit pour mettre en évidence tous les défauts et avantages importants du schéma qui ont trait à la convergence des solutions approchées obtenues avec son aide.

Les équations modèles pour les systèmes d'équations et les équations d'ordre supérieur se construisent de la même façon.

Dans la pratique, la solution du problème de Cauchy pour des équations différentielles ordinaires sans singularités se calcule d'habitude selon un ou deux schémas éprouvés assez universels pour lesquels il existe des programmes standard exploités sur les ordinateurs modernes.

Si on a à résoudre des problèmes d'un type particulier avec une précision très élevée, on s'adresse à des schémas spéciaux à ces problèmes, qui sont nombreux mais qui le cèdent aux schémas universels pour ce qui est des problèmes de type différent.

CHAPITRE 6

SCHÉMAS AUX DIFFÉRENCES USUELS

§19. Schémas de Runge-Kutta et d'Adams

Décrivons certains schémas aux différences usuels de la résolution du problème de Cauchy pour l'équation différentielle du premier ordre

$$\left.\begin{aligned} \frac{du}{dx}-G(x, u) &= 0, \quad 0 \leqslant x \leqslant 1, \\ u(0) &= a. \end{aligned}\right\} \tag{1}$$

A la fin du présent paragraphe nous les généraliserons à des systèmes d'équations différentielles du premier ordre auxquels se réduit le cas général d'équations et de systèmes d'ordre quelconque.

Discrétisons le segment $0 \leqslant x \leqslant 1$:

$$0 = x_0 < x_1 < x_2 < \ldots < x_{N-1} < x_N = 1, \quad x_n = nh, \quad h = 1/N,$$

et construisons des schémas aux différences pour obtenir approximativement la table $[u]_h$ de valeurs de la solution $u(x)$ sur le réseau choisi.

Le schéma usuel le plus simple nous est déjà familier : c'est le schéma d'Euler

$$L_h u^{(h)} \equiv \left\{\begin{aligned} \frac{u_{n+1}-u_n}{h}-G(x_n, u_n) &= 0, \quad n = 0, 1, \ldots, N-1, \\ u_0 &= a, \end{aligned}\right. \tag{2}$$

approchant (et exact) à l'ordre 1.

Le calcul d'après ce schéma a un sens géométrique simple. Si u_n est déjà trouvé, le calcul de

$$u_{n+1} = u_n+hG(x_n, u_n)$$

équivaut à un déplacement du point (x_n, u_n) au point (x_{n+1}, u_{n+1}) du plan Oxu suivant la tangente à la courbe intégrale $u = u(x)$ passant par le point (x_n, u_n) de l'équation différentielle $u' = G(x, u)$.

Parmi les schémas d'ordre d'approximation supérieur, on utilise surtout diverses variantes de Runge-Kutta et d'Adams que nous allons décrire et comparer.

1. Schémas de Runge-Kutta. La valeur u_n de la solution approchée au point x_n étant trouvée, on demande de calculer u_{n+1} au point $x_{n+1} = x_n+h$.

Donnons-nous un l entier et écrivons les expressions

$$\begin{aligned} k_1 &= G(x_n, u_n), \\ k_2 &= G(x_n+\alpha h, u_n+\alpha h k_1), \\ k_3 &= G(x_n+\beta h, u_n+\beta h k_2), \\ &\dots\dots\dots\dots\dots\dots \\ k_l &= G(x_n+\gamma h, u_n+\gamma h k_{l-1}). \end{aligned}$$

Posons

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1}-u_n}{h} - (p_1 k_1 + \dots + p_l k_l) = 0, & n = 0, 1, \dots, N-1, \\ u_0 = a. \end{cases}$$

l étant donné, choisissons les coefficients $\alpha, \beta, \dots, \gamma, p_1, p_2, \dots, p_l$ de façon que l'approximation soit d'ordre le plus élevé possible. Connaissant u_n, on peut calculer $k_1, \dots, k_l$, puis

$$u_{n+1} = u_n + h(p_1 k_1 + \dots + p_l k_l).$$

Le schéma d'Euler ($l = 1$) est le plus simple des schémas de Runge-Kutta.

Le schéma de Runge-Kutta

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1}-u_n}{h} - \dfrac{h}{6}(k_1+2k_2+2k_3+k_4) = 0, \\ n = 0, 1, \dots, N-1, \\ u_0 = a, \end{cases} \tag{3}$$

où

$$\begin{aligned} k_1 &= G(x_n, u_n), \\ k_2 &= G\left(x_n+\frac{h}{2}, u_n+\frac{k_1 h}{2}\right), \\ k_3 &= G\left(x_n+\frac{h}{2}, u_n+\frac{k_2 h}{2}\right), \\ k_4 &= G(x_n+h, u_n+k_3 h), \end{aligned}$$

approche à l'ordre quatre.

L'approximation par le schéma de Runge-Kutta

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1}-u_n}{h} - \left[\dfrac{2\alpha-1}{2\alpha} k_1 + \dfrac{1}{2\alpha} k_2\right] = 0, \\ n = 0, 1, \dots, N-1, \\ u_0 = a, \end{cases} \tag{4}$$

où

$$k_1 = G(x_n, u_n), \quad k_2 = G(x_n+\alpha h, u_n+\alpha h k_1),$$

pour tout α fixe est d'ordre 2.

Démontrons la dernière assertion. (Dans le cas du schéma (3) la démonstration est analogue mais plus laborieuse.)

La solution $u(x)$ de l'équation $u' = G(x, u)$ vérifie les identités

$$\frac{du}{dx} \equiv G(x, u(x)),$$

$$\frac{d^2u}{dx^2} \equiv \frac{d}{dx} G(x, u) = \frac{\partial G}{\partial x} + \frac{\partial G}{\partial u} G.$$

La formule de Taylor

$$\frac{u(x_n+h)-u(x_n)}{h} = u'(x_n)+\frac{h}{2} u''(x_n)+O(h^2)$$

pour $u(x)$ entraîne donc

$$\frac{u(x_{n+1})-u(x_n)}{h} - \left[G+\frac{h}{2}\left(\frac{\partial G}{\partial x}+\frac{\partial G}{\partial u} G\right)\right]_{\substack{x=x_n \\ u=u(x_n)}} = O(h^2). \tag{5}$$

Mais en représentant la fonction de deux variables par son développement taylorien suivant h et en retenant les termes du premier degré, on a

$$\frac{2\alpha-1}{2\alpha} k_1+\frac{1}{2\alpha} k_2\Big|_{\substack{x=x_n \\ u=u(x_n)}} = \frac{2\alpha-1}{2\alpha} G+\frac{1}{2\alpha} G(x+\alpha h,\ u+\alpha hG)\Big|_{\substack{x=x_n \\ u=u(x_n)}} =$$

$$= \frac{2\alpha-1}{2\alpha} G+\frac{1}{2\alpha}\left[G+\frac{\partial G}{\partial x}\alpha h+\frac{\partial G}{\partial u}\alpha hG+O(h^2)\right]_{\substack{x=x_n \\ u=u(x_n)}} =$$

$$= G+\frac{h}{2}\left(\frac{\partial G}{\partial x}+\frac{\partial G}{\partial u} G\right)\Big|_{\substack{x=x_n \\ u=u(x_n)}} + O(h^2). \tag{6}$$

En substituant à u_n et u_{n+1} du premier membre de (4) les valeurs respectives $u(x_n)$ et $u(x_{n+1})$ de la solution $u(x)$, on obtient donc une expression coïncidant avec le premier membre de (5) à $O(h^2)$ près. C'est, par conséquent, une expression du second ordre en h. La valeur $u_0 = a$ étant donnée de façon exacte, cela achève la démonstration de notre assertion.

Pour trouver u_{n+1} par le schéma de Runge-Kutta avec u_n connu les valeurs de la fonction $G(x, u)$ sont calculées l fois et ne sont plus utilisées.

2. Schémas d'Adams. Dans les schémas d'Adams dont nous allons décrire une variante, on trouve chaque valeur suivante u_{n+1} en calculant une valeur supplémentaire de $G(x, u)$ en un seul point quel que soit l'ordre d'approximation. Il faut de plus effectuer des soustractions et additions peu nombreuses qui coûtent beaucoup moins en temps que chaque calcul d'une fonction $G(x, u)$ tant soit peu compliquée.

Introduisons les notations

$$\nabla f_n = f_n - f_{n-1},$$

$$\nabla^2 f_n = \nabla(\nabla f_n) = \nabla f_n - \nabla f_{n-1} = f_n - 2f_{n-1} + f_{n-2},$$

$$\nabla^3 f_n = \nabla\nabla^2 f_n = f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3}$$

et posons $G_n = G(x_n, u_n)$. Ecrivons plusieurs équations aux différences utilisées dans les schémas d'Adams pour rechercher u_{n+1} si $u_n, u_{n-1}, \ldots$ le sont déjà :

$$\frac{u_{n+1}-u_n}{h} - G_n = 0, \quad n = 0, 1, \ldots, N-1, \tag{7}$$

$$\frac{u_{n+1}-u_n}{h} - G_n - \frac{1}{2}\nabla G_n = 0, \quad n = 1, 2, \ldots, N-1, \tag{8}$$

$$\frac{u_{n+1}-u_n}{h} - G_n - \frac{1}{2}\nabla G_n - \frac{5}{12}\nabla^2 G_n = 0, \quad n = 2, 3, \ldots, N-1, \tag{9}$$

$$\frac{u_{n+1}-u_n}{h} - G_n - \frac{1}{2}\nabla G_n - \frac{5}{12}\nabla^2 G_n - \frac{3}{8}\nabla^3 G_n = 0, \quad n = 3, 4, \ldots, N-1. \tag{10}$$

La première est une équation aux différences d'Euler. Le report des valeurs $u((n+1)h)$, $u(nh)$, ... de la solution exacte $u(x)$ dans les premiers membres des équations (7) à (10) où elles se substituent à $u_{n+1}, u_n, u_{n-1}, \ldots$ y fait naître des résidus en h, h^2, h^3 et h^4 respectivement.

Les formules d'Adams s'obtiennent comme suit. Soit $u(x)$ solution de l'équation

$$\frac{du}{dx} = G(x, u).$$

Notons

$$G(x, u(x)) \equiv F(x).$$

Alors

$$u(x_n+h) - u(x_n) = \int_{x_n}^{x_n+h} u'\,dx = \int_{x_n}^{x_n+h} F(x)\,dx.$$

La théorie de l'interpolation des fonctions apprend qu'il existe un polynôme $P_k(x, F)$, et un seul, de degré k au plus qui prend aux $k+1$ points $x_n, x_{n-1}, \ldots, x_{n-k}$ des valeurs données respectives $F(x_n), F(x_{n-1}), \ldots, F(x_{n-k})$. Si $F(x)$ est suffisamment régulière, $P_k(x, F)$ s'en écarte sur le segment $x_n \leqslant x \leqslant x_n + h$ d'une quantité de l'ordre de h^{k+1} si bien que

$$\max |P_k(x, F) - F(x)| = O(h^{k+1}). \tag{11}$$

La formule aux différences d'Adams s'écrit

$$\frac{u_{n+1}-u_n}{h} - \frac{1}{h}\int_{x_n}^{x_n+h} P_k(x, F)\,dx = 0. \tag{12}$$

En remplaçant au premier membre

$$u_n, u_{n+1}, G(x_{n-s}, u_{n-s})$$

par

$$u(x_n), u(x_{n+1}), G(x_{n-s}, u(x_{n-s})),$$

on obtient un résidu en h^{k+1} :

$$\left|\frac{u(x_n+h)-u(x_n)}{h}-\frac{1}{h}\int\limits_{x_n}^{x_n+h} P_k(x, F)\,dx\right| =$$

$$= \left|\left[\frac{u(x_n+h)-u(x_n)}{h}-\frac{1}{h}\int\limits_{x_n}^{x_n+h} F(x)\,dx\right]+\right.$$

$$\left.+\frac{1}{h}\int\limits_{x_n}^{x_n+h} [F(x)-P_k(x, F)]\,dx\right| \leqslant 0+\max|F(x)-P_k(x, F)| = O(h^{k+1}).$$

Pour $k = 0$, le polynôme d'interpolation

$$P_0(x, F) = G(x_n, u_n) = \text{const},$$

et la formule (12) change en (7).

Pour $k = 1$

$$P_1(x, F) = \frac{1}{h}[(x-x_{n-1})G_n-(x-x_n)G_{n-1}].$$

Ensuite,

$$\frac{1}{h}\int\limits_{x_n}^{x_n+h} P_1(x, F)\,dx = \frac{1}{h^2}\,\frac{(x-x_{n-1})^2}{2}\Bigg|_{x_n}^{x_n+h} G_n-$$

$$-\frac{1}{h^2}\,\frac{(x-x_n)^2}{2}\Bigg|_{x_n}^{x_n+h} G_{n-1} = \frac{1}{h^2}\left(\frac{4h^2}{2}-\frac{h^2}{2}\right)G_n-\frac{1}{h^2}\,\frac{h^2}{2}\,G_{n-1} =$$

$$= G_n+\frac{1}{2}\nabla G_n.$$

La formule (12) se transforme donc en (8). De même, pour $k = 2$ et $k = 3$, l'égalité (12) devient respectivement (9) et (10).

Le calcul par le schéma (7) exige la connaissance de seul $u_0 = a$ et par le schéma (8) celle de $u_0 = a$ et de u_1. Le schéma (9) démarre à condition de connaître u_0, u_1, u_2 et le schéma (10) avec u_0, u_1, u_2 et u_3 trouvés. Ces valeurs s'obtiennent par le schéma de Runge-Kutta, par la variante d'Euler à pas faible et moyennant le développement de la solution en série de puissances au voisinage du point $x = 0$. Une mise en route non standard constitue l'un des défauts des schémas d'Adams devant ceux de Runge-Kutta et l'un de leurs avantages est, comme nous l'avons déjà dit, de calculer, pour obtenir u_{n+1}, une seule valeur de $G_n = G(x_n, u_n)$ en plus de G_s, $\nabla G_s, \ldots, \nabla^k G_s$ trouvés en recherchant $u_n, u_{n-1}, \ldots$ et d'effectuer un certain nombre de soustractions donnant $\nabla G_n, \ldots, \nabla^k G_n$.

Ainsi, les méthodes d'Adams l'emportent sur celles de Runge-Kutta en ce qui concerne l'effort de calcul par pas. Leurs inconvénients principaux sont un démarrage non standard, l'impossibilité (sans alourdir les formules) de modifier, à partir d'un point x_n, le pas h, $x_{n+1} = x_n+h$. Cette dernière circonstance se fait particulièrement sentir si la solution et ses dérivées varient vite sur certains secteurs et lentement sur d'autres.

Si un fait pareil est mis en évidence au cours du calcul, le schéma de Runge-Kutta peut, par exemple en exécutant un sous-programme, diminuer ou augmenter automatiquement le pas sur les secteurs régulieurs afin d'éviter un travail superflu.

La stratégie la plus rationnelle est visiblement la mise en œuvre des deux techniques, celle de Runge-Kutta et celle d'Adams, avec passage automatique de l'une à l'autre en cours de route. Cela étant, on démarrera le calcul par un schéma de Runge-Kutta. Le programme doit prévoir un choix automatique du pas assurant la précision désirée, et ce choix doit être « conservatif » dans un certain sens : le programme ne fait modifier le pas que s'il le faut d'une façon « assez sensible ». Si, en calculant plusieurs valeurs successives de u_n par un schéma de Runge-Kutta, le pas ne varie pas, il y a intérêt à passer automatiquement à un schéma d'Adams plus économique. Dès qu'une nécessité apparaît de changer le pas, on retourne au schéma de Runge-Kutta, et ainsi de suite.

On vérifie le choix du pas par un calcul parallèle avec un pas donné deux fois moindre. Dans les limites de la précision voulue les solutions doivent coïncider. Dans le cas contraire, un pas plus petit s'impose. Il faut également prévoir un test sur la possibilité d'augmenter le pas.

3. Remarques sur la stabilité. Dans le cas du problème linéaire $u' + Au = 0$ à coefficient A constant, après élimination de $k_1, k_2, \ldots$, les schémas de Runge-Kutta sont du premier ordre

$$u_{n+1} - a(h)u_n = 0.$$

La racine de l'équation caractéristique $\lambda - a(h) = 0$ vaut $\lambda = a(h)$.

Si $u_n = u(x_n)$, la valeur u_{n+1} coïncide avec la solution exacte $u(x_n+h)$ à h^{p+1} près, p étant l'ordre d'approximation. Puisque

$$u(x_n+h) = u(x_n)e^{-Ah_n} = u(x_n)\left(1 - Ah + \frac{A^2h^2}{2} - \ldots\right)$$

et

$$u_{n+1} = a(h)u_n,$$

on a

$$\lambda = a(h) = e^{-Ah} + O(h^{p+1}).$$

Donc

$$|\lambda(h)| < 1 + ch.$$

Les puissances $\lambda^n(h)$ se comportent de façon « régulière » : elles croissent si $A < 0$ et la solution de l'équation différentielle croît. Elles décroissent si $A > 0$ et la solution e^{-Ax} décroît.

Pour le schéma d'Adams (8)

$$\frac{u_{n+1} - u_n}{h} + Au_n + \frac{A}{2}(u_n - u_{n-1}) = 0 \tag{13}$$

l'équation caractéristique est de la forme

$$\lambda^2-\left(1-\frac{3Ah}{2}\right)\lambda-\frac{Ah}{2}=0.$$

D'où

$$\lambda=\frac{1}{2}\left(1-\frac{3Ah}{2}\right)\pm\frac{1}{2}\sqrt{\left(1-\frac{3Ah}{2}\right)^2+2Ah},$$
$$\lambda_1=1-Ah+O(h^2),$$
$$\lambda_2=O(h).$$

Ainsi, la solution $u_n=\lambda_1^n$ se comporte comme $u(x_n)=e^{-Anh}$ avec la diminution de h, et la solution « parasite » λ_2^n due au choix de l'équation aux différences d'ordre 2 tend vers zéro vu que $|\lambda_2|=O(h)$ et n'influe pas sur la stabilité.

Le lecteur ferait bien de comparer le schéma (13) avec le schéma d'ordre 2 (2) du § 17 :

$$\frac{u_{n+1}-u_{n-1}}{2h}+Au_n=0.$$

Pour ce dernier

$$\lambda_1=1-Ah+\frac{A^2h^2}{2}+O(h^3),\qquad \lambda_2=-1-Ah+O(h^2).$$

Pour A positif, la racine « parasite » λ_2 est supérieure en module à λ_1, d'où une valeur importante de la constante dans l'estimation de la stabilité pour ce schéma et son impuissance pour A grands établie au § 17.

4. Généralisation à des systèmes d'équations. Tous les schémas décrits utilisés pour la résolution numérique du problème de Cauchy pour l'équation différentielle d'ordre 1 (1) s'étendent automatiquement à des systèmes d'équations du premier ordre. Pour cela, dans l'écriture (1)

$$\left.\begin{aligned}\frac{du}{dx}-G(x,u)&=0,\\ u(0)&=a\end{aligned}\right\}$$

$u(x)=\bar{u}(x)$, $G(x,u)=\bar{G}(x,\bar{u})$ sont des fonctions vectorielles et $a=\bar{a}$ est un vecteur donné. Dans ces conditions les schémas de Runge-Kutta (3), (4) et les schémas d'Adams (7) à (10) ont toujours un sens et restent valides.

Ecrivons par exemple le système d'équations

$$\left.\begin{aligned}\frac{dv}{dx}-(x+v^2+\sin w)&=0,\\ \frac{dw}{dx}+xvw&=0,\end{aligned}\right\}$$
$$\left.\begin{aligned}v(0)&=a_1,\\ w(0)&=a_2\end{aligned}\right\}$$

sous la forme

$$\left.\begin{aligned}\frac{d\bar{u}}{dx}-G(x,\bar{u}) &= 0,\\ \bar{u}(0) &= \bar{a},\end{aligned}\right\}$$

en posant

$$\bar{u}(x) = \begin{pmatrix} v(x)\\ w(x)\end{pmatrix},$$

$$\bar{G}(x,\bar{u}) = \begin{pmatrix} x+v^2+\sin w\\ -xvw\end{pmatrix},$$

$$\bar{a} = \begin{pmatrix} a_1\\ a_2\end{pmatrix}.$$

La formule pour $\bar{u}_{n+1}$ dans le schéma d'Euler

$$\bar{u}_{n+1} = \bar{u}_n+h\bar{G}(x_n,\bar{u}_n)$$

s'écrira en détaillant

$$\left.\begin{aligned}v_{n+1} &= v_n+h(x_n+v_n^2+\sin w_n),\\ w_{n+1} &= w_n+h(-x_nv_nw_n).\end{aligned}\right\}$$

Tous les raisonnements de la page 145 sur l'ordre d'approximation restent également en vigueur. Cela étant, on entend par la dérivée $\frac{\partial G}{\partial u}$ du vecteur $G(G_1, \ldots, G_k)$ par rapport au vecteur $u(u_1, \ldots, u_k)$ de la formule (6) la matrice

$$\begin{pmatrix} \frac{\partial G_1}{\partial u_1} & \cdots & \frac{\partial G_1}{\partial u_k}\\ \cdots & \cdots & \cdots\\ \frac{\partial G_k}{\partial u_1} & \cdots & \frac{\partial G_k}{\partial u_k}\end{pmatrix}.$$

Un système arbitraire d'équations différentielles résolues en les dérivées d'ordre le plus élevé se réduit au système d'équations du premier ordre

$$\frac{d\bar{u}}{dx} = \bar{G}(x,\bar{u})$$

par substitution des fonctions inconnues. L'exemple suivant l'illustre clairement. Le système

$$\left.\begin{aligned}\frac{d^2v}{dx^2}+\sin(xv'+v^2+w) &= 0,\\ \frac{dw}{dx}+\sqrt{x^2+v^2+(v')^2+w^2} &= 0,\\ v(0) &= a,\\ v'(0) &= b,\\ w(0) &= c\end{aligned}\right\}$$

se ramène à la forme nécessaire à condition de poser

$$u_1(x) = v(x),$$
$$u_2(x) = \frac{dv}{dx},$$
$$u_3(x) = w(x).$$

On obtient

$$\left.\begin{aligned} \frac{du_1}{dx} - u_2 &= 0, \\ \frac{du_2}{dx} + \sin(xu_2 + u_1^2 + u_3) &= 0, \\ \frac{du_3}{dx} + \sqrt{x^2 + u_1^2 + u_2^2 + u_3^2} &= 0, \\ u_1(0) &= a, \\ u_2(0) &= b, \\ u_3(0) &= c. \end{aligned}\right\}$$

REMARQUE. On a élaboré des schémas aux différences du type Runge-Kutta qui s'appliquent directement à des équations d'ordre 2 sans qu'il soit nécessaire de réduire préalablement celles-ci à des systèmes d'ordre 1.

§20. Méthodes de résolution des problèmes aux limites

Soit le problème aux limites

$$\left.\begin{aligned} y'' &= f(x, y, y'), \quad 0 \leqslant x \leqslant 1, \\ y(0) &= Y_0, \qquad\qquad y(1) = Y_1, \end{aligned}\right\} \tag{1}$$

avec les conditions aux deux extrémités du segment $0 \leqslant x \leqslant 1$ où on cherche une solution $y = y(x)$. Cet exemple nous permettra d'exposer de façon schématique certains procédés de résolution numérique des problèmes aux limites.

1. Méthode du tir. Au §19 nous avons cité des modes opératoires commodes pour résoudre numériquement le problème de Cauchy, i. e. un problème de la forme

$$\left.\begin{aligned} y'' &= f(x, y, y'), \quad 0 \leqslant x \leqslant 1, \\ y(0) &= Y_0, \quad \left.\frac{dy}{dx}\right|_{x=0} = \operatorname{tg}\alpha, \end{aligned}\right\} \tag{2}$$

Y_0 étant l'ordonnée du point $(0, Y_0)$, origine de la courbe intégrale, et α l'angle de la courbe intégrale et de l'axe Ox en ce point (fig. 7,*a*). Pour Y_0 fixe, la solution du problème (2) s'obtient comme $y = y(x, \alpha)$. Quand $x = 1$, la solution $y(x, \alpha)$ ne dépend que de α :

$$y(x, \alpha)\,|_{x=1} = y(1, \alpha).$$

La remarque faite sur la solution du problème de Cauchy (2) aidant, nous énoncerons comme suit le problème (1) : trouver un angle $\alpha = \alpha^*$ tel que la courbe intégrale faisant en son point origine $(0, Y_0)$ un angle α avec l'axe des abscisses tombe au point $(1, Y_1)$:

$$y(1, \alpha) = Y_1. \tag{3}$$

La solution du problème (2) coïncide pour ce $\alpha = \alpha^*$ avec la solution cherchée de (1). On est donc ramené à résoudre l'équation (3) (fig. 7,*b*), qui est

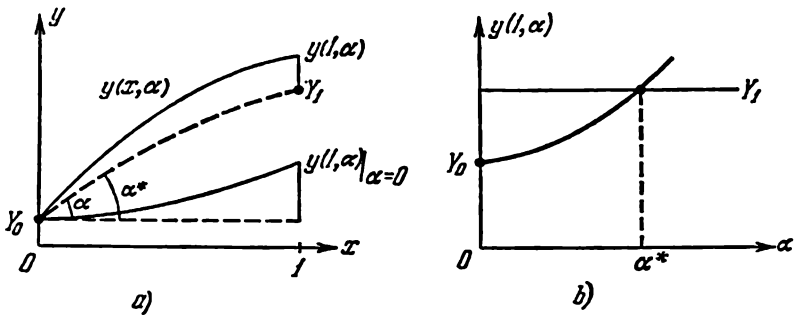


Fig. 7

de la forme $F(\alpha) = 0$, avec $F(\alpha) = y(1, \alpha) - Y_1$. Sa seule différence par rapport aux équations usuelles est que la fonction $F(\alpha)$ est donnée non pas par une expression analytique, mais à l'aide de l'algorithme de résolution du problème (2).

Le fait de ramener la résolution du problème aux limites (1) à celle du problème de Cauchy (2) est justement à la base de la méthode du tir.

L'équation (3) peut se résoudre par le procédé de bipartition du segment, par la méthode de la tangente (ou de Newton), la méthode des parties proportionnelles, etc. Dans la première méthode par exemple on se donne α_0 et α_1 tels que les différences

$$y(1, \alpha_0) - Y_1 \quad \text{et} \quad y(1, \alpha_1) - Y_1$$

soient de signes opposés. On pose ensuite

$$\alpha_2 = \frac{\alpha_0 + \alpha_1}{2}.$$

On calcule $y(1, \alpha_2)$. On trouve α_3 par l'une des formules

$$\alpha_3 = \frac{\alpha_0 + \alpha_2}{2} \quad \text{ou} \quad \alpha_3 = \frac{\alpha_1 + \alpha_2}{2}$$

selon que les différences

$$y(1, \alpha_1) - Y_1 \quad \text{et} \quad y(1, \alpha_2) - Y_1$$

sont de même signe ou non. Puis on calcule $y(1, \alpha_3)$ et on continue tant que n'est pas atteinte la précision désirée, $|y(1, \alpha_n) - Y_1| < \varepsilon$.

En procédant par parties proportionnelles on se donne α_0 et α_1, et on calcule les α_i suivants par la formule de récurrence

$$\alpha_{n+1} = \alpha_n - \frac{F(\alpha_n)}{F(\alpha_n) - F(\alpha_{n-1})}(\alpha_n - \alpha_{n-1}), \quad n = 1, 2, \ldots$$

La méthode du tir qui réduit la résolution du problème aux limites (1) à celle du problème de Cauchy (2) s'avère puissante si la solution $y(x, \alpha)$ ne dépend pas « trop » de α. Dans le cas contraire, le procédé est numériquement instable même pour une dépendance « modérée » de la solution du problème (1) par rapport aux données initiales.

Expliquons ce que nous voulons dire en guillemetant certains mots et considérons à cette fin le problème aux limites linéaire

$$\left.\begin{aligned} y'' - a^2 y &= 0, \quad 0 \leqslant x \leqslant 1, \\ y(0) &= Y_0, \quad y(1) = Y_1, \end{aligned}\right\} \tag{1'}$$

a^2 étant constant. Ecrivons sa solution :

$$y(x) = \frac{e^{-ax} - e^{-a(2-x)}}{1 - e^{-2a}} Y_0 + \frac{e^{-a(1-x)} - e^{-a(1+x)}}{1 - e^{-2a}} Y_1.$$

Pour a croissant, les coefficients de Y_0 et Y_1 restent des fonctions bornées sur le segment $0 \leqslant x \leqslant 1$, et, quels que soient $a > 0$, ils sont au plus égaux à l'unité. C'est pourqoui en commettant de petites erreurs à la donnée de Y_0 et Y_1 on induit des erreurs tout aussi faibles sur la solution. Considérons maintenant le problème de Cauchy

$$\left.\begin{aligned} y'' - a^2 y &= 0, \quad 0 \leqslant x \leqslant 1, \\ y(0) &= Y_0, \quad y'(0) = \operatorname{tg} \alpha \end{aligned}\right\} \tag{2'}$$

dont la solution est de la forme

$$y(x) = \frac{aY_0 + \operatorname{tg} \alpha}{2a} e^{ax} + \frac{aY_0 - \operatorname{tg} \alpha}{2a} e^{-ax}.$$

Si $\operatorname{tg} \alpha$ donné est faux de ε, la valeur de la solution acquiert pour $x = 1$ l'accroissement

$$\Delta y(1) = \frac{\varepsilon}{2a} e^{a} - \frac{\varepsilon}{2a} e^{-a}. \tag{4}$$

Pour a grands, le dernier terme de l'égalité (4) est négligeable, mais le coefficient de ε du premier terme de la différence devient important. Tout en étant formellement acceptable, la méthode du tir est donc impuissante à résoudre (1') pour a grands. Cela fait penser à nos raisonnements du n° 2 du § 5 où nous avons donné l'exemple d'un algorithme numériquement instable de résolution d'un problème aux limites aux différences.

2. Méthode du balayage. On résout le problème aux limites

$$\left.\begin{aligned} y'' - p(x) y &= f(x), \quad 0 \leqslant x \leqslant 1, \\ y(0) &= Y_0, \quad y(1) = Y_1 \end{aligned}\right\}$$

pour $p(x) \gg 1$ en se servant du schéma aux différences

$$\left.\begin{aligned} &\frac{y_{m+1}-2y_m+y_{m-1}}{h^2}-p(x_m)y_m = f(x_m), \\ &0 < m < M, \quad Mh = 1, \\ &y_0 = Y_0, \quad y_M = Y_1 \end{aligned}\right\}$$

et en appliquant au problème aux différences le procédé par balayage. Les conditions de validité du balayage pour $p(x) > 0$ sont remplies (le lecteur s'en rendra aisément compte).

3. Méthode de Newton. La méthode du tir pour un problème aux limites bien posé peut s'avérer, nous l'avons vu, instable sur le plan numérique, donc impraticable, et la méthode du balayage ne s'applique, même formellement, que dans le cas de problèmes linéaires.

L'idée de la méthode de Newton consiste à réduire la résolution d'un problème non linéaire à celle d'une suite de problèmes linéaires. Etant donné une fonction $y_0(x)$ vérifiant les conditions aux limites (1) et approchant grossièrement la solution cherchée $y(x)$, posons

$$y(x) = y_0(x)+v(x), \tag{5}$$

où $v(x)$ est une correction à l'approximation d'ordre zéro $y_0(x)$. Portons (5) dans l'équation (1) et linéarisons le problème à l'aide de

$$y''(x) = y_0''(x)+v''(x),$$

$$f(x, y_0+v, y_0'+v') = f(x, y_0, y_0')+\frac{\partial f(x, y_0, y_0')}{\partial y}v+\frac{\partial f(x, y_0, y_0')}{\partial y'}v'+ \\ +O(v^2+|v'|^2).$$

Rejetons le reste $O(v^2+|v'|^2)$, il vient un problème linéaire pour la correction $\bar{v}(x)$:

$$\left.\begin{aligned} \bar{v}'' &= p(x)\bar{v}'+q(x)\bar{v}+\varphi(x), \\ \bar{v}(0) &= \bar{v}(1) = 0, \end{aligned}\right\} \tag{6}$$

où

$$p(x) = \frac{\partial f(x, y_0, y_0')}{\partial y'}, \quad q(x) = \frac{\partial f(x, y_0, y_0')}{\partial y}, \\ \varphi(x) = f(x, y_0, y_0')-y_0''.$$

En résolvant (6) analytiquement ou par une méthode numérique, trouvons approximativement la correction $\bar{v}$ et prenons

$$y_1 \equiv y_0(x)+\bar{v}$$

pour l'approximation suivante.

La procédure décrite reste valable pour le problème non linéaire intervenant dans l'approximation du problème (1).

TROISIÈME PARTIE

SCHÉMAS AUX DIFFÉRENCES POUR LES ÉQUATIONS AUX DÉRIVÉES PARTIELLES. NOTIONS FONDAMENTALES

En parlant des schémas aux différences pour les équations différentielles ordinaires nous avons défini la convergence, l'approximation et la stabilité. Un théorème a été démontré aux termes duquel, un problème aux limites aux différences approchant un problème différentiel et étant stable, sa solution converge avec le resserrement du réseau vers celle du problème différentiel. Ce théorème suggère les procédés de construction de schémas aux différences convergents utilisés pour la résolution numérique des problèmes aux limites différentiels : il faut construire des schémas aux différences approximants et en sélectionner stables.

Les définitions de la convergence, de l'approximation, de la stabilité et le théorème reliant ces trois notions sont de caractère général. Ils ont un sens pour n'importe quelle équation fonctionnelle. Plus haut nous les avons illustrés sur des schémas aux différences pour les équations différentielles ordinaires et pour l'équation intégrale. Nous nous proposons maintenant de donner plusieurs techniques principales de construction de schémas aux différences et de vérifier leur stabilité sur les exemples de schémas pour les équations aux dérivées partielles. Nombre de circonstances nouvelles importantes se feront jour dont les principales sont la diversité des réseaux et des procédés d'approximation, l'instabilité de la plupart des schémas approximants choisis au hasard, les difficultés que présentent l'étude de la stabilité et le calcul de solutions des problèmes aux limites aux différences (ce dernier exige un effort mathématique spécial).

CHAPITRE 7

PROCÉDÉS SIMPLES DE CONSTRUCTION ET D'ÉTUDE DE SCHÉMAS AUX DIFFÉRENCES

§21. Rappels et illustration des définitions principales

1. Convergence. On demande de calculer approximativement la solution u du problème aux limites différentiel

$$Lu = f \tag{1}$$

posé dans un domaine D de frontière Γ. On choisit à cette fin un ensemble discret de points D_h (un réseau) appartenant à $D+\Gamma$, on introduit un espace vectoriel normé U_h de fonctions définies sur D_h et on établit une correspondance entre la solution u et une fonction $[u]_h \in U_h$, la table cherchée de u. Pour approcher la table $[u]_h$, que nous avons convenu de prendre pour solution exacte du problème (1), il faut, à partir de (1), former un système d'équations

$$L_h u^{(h)} = f^{(h)} \tag{2}$$

en $u^{(h)}$ de U_h tel qu'on ait convergence

$$\| [u]_h - u^{(h)} \|_{U_h} \to 0 \quad \text{pour } h \to 0. \tag{3}$$

Si la solution du problème aux limites aux différences (2) vérifie l'inégalité

$$\| [u]_h - u^{(h)} \|_{U_h} \leqslant Ch^k,$$

la convergence est dite d'ordre k en h.

La construction du schéma aux différences convergent (2) est partagée en deux étapes : la construction du schéma (2) approchant le problème (1) pour la solution u de celui-ci et la vérification de la stabilité du schéma (2).

2. Approximation. Rappelons la définition de l'approximation. Pour que cette notion ait un sens, on introduit une norme dans l'espace F_h dont fait partie le second membre $f^{(h)}$ de l'équation (2). Le problème aux différences (2) approche par définition le problème (1) pour la solution u si dans l'égalité

$$L_h[u]_h = f^{(h)} + \delta f^{(h)}$$

le résidu $\delta f^{(h)}$ dû à la substitution de $[u]_h$ dans le problème aux limites aux différences (2) tend vers zéro avec h :

$$\| \delta f^{(h)} \|_{F_h} = \| L_h[u]_h - f^{(h)} \|_{F_h} \to 0.$$

Si

$$\| \delta f^{(h)} \|_{F_h} \leqslant Ch^k,$$

avec C indépendant de h, l'approximation est d'ordre k en h.

Construisons, par exemple, pour le problème de Cauchy

$$\left.\begin{aligned} \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} &= \varphi(x, t), \quad -\infty < x < \infty, \quad 0 \leqslant t \leqslant T, \\ u(x, 0) &= \psi(x), \quad -\infty < x < \infty \end{aligned}\right\} \tag{4}$$

un schéma aux différences approximant. Le problème (4) prend la forme (1) en posant

$$Lu \equiv \begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial u}{\partial x}, & -\infty < x < \infty, \quad 0 \leqslant t \leqslant T, \\ u(x, 0), & -\infty < x < \infty, \end{cases}$$

$$f = \begin{cases} \varphi(x, t), & -\infty < x < \infty, \quad 0 \leqslant t \leqslant T, \\ \psi(x), & -\infty < x < \infty. \end{cases}$$

Le réseau D_h sera l'ensemble de points d'intersection des droites

$$x = mh, \quad t = n\tau, \quad m = 0, \pm 1, \ldots ; \quad n = 0, 1, \ldots, [T/\tau],$$

où $h > 0$, $\tau > 0$ sont des nombres et $[T/\tau]$ la partie entière de T/τ (fig. 8).

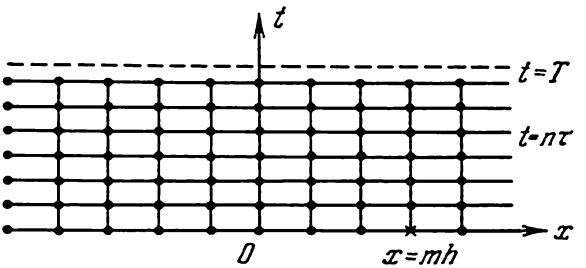


Fig. 8

Supposons qu'il y a entre le pas h et le pas τ la relation $\tau = rh$, r étant constant, de sorte que D_h est fonction du seul paramètre h. La fonction discrète cherchée est la table $[u]_h = \{u(mh, n\tau)\}$ de valeurs de la solution $u(x, t)$ du problème (4) aux points du réseau D_h.

Construisons le schéma aux différences (2) approchant le problème (4). Les valeurs de u^h au point $(x_m, t_n) = (mh, n\tau)$ du réseau D_h seront désignées par u_m^n. Le schéma (2) s'obtient en approchant les dérivées $\partial u/\partial t$ et $\partial u/\partial x$ par les rapports aux différences

$$\left.\begin{aligned} \left.\frac{\partial u}{\partial t}\right|_{x,t} &\approx \frac{u(x, t+\tau) - u(x, t)}{\tau}, \\ \left.\frac{\partial u}{\partial x}\right|_{x,t} &\approx \frac{u(x+h, t) - u(x, t)}{h}. \end{aligned}\right\} \tag{4'}$$

Ce schéma est de la forme

$$\left.\begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = \varphi(mh, n\tau), \\ &m = 0, \pm 1, \ldots ; \quad n = 0, 1, \ldots, [T/\tau] - 1, \\ &u_m^0 = \psi(mh), \quad m = 0, \pm 1, \ldots \end{aligned}\right\} \tag{5}$$

L'opérateur L_h et le second membre $f^{(h)}$ pour le schéma (5) sont donnés par les égalités respectives

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^n - u_m^n}{h}, \\ \qquad m = 0, \pm 1, \ldots ; \quad n = 0, 1, \ldots, [T/\tau] - 1, \\ u_m^0, \quad m = 0, \pm 1, \ldots, \end{cases}$$

$$f^{(h)} = \begin{cases} \varphi(mh, n\tau), & m = 0, \pm 1, \ldots, \quad n = 0, 1, \ldots, [T/\tau] - 1, \\ \psi(mh), & m = 0, \pm 1, \ldots \end{cases}$$

On voit donc que $f^{(h)}$ est un couple de fonctions discrètes $\varphi(mh, n\tau)$ et $\psi(mh)$ dont l'une est définie sur le réseau bidimensionnel

$$(x_m, t_n) = (mh, n\tau), \quad m = 0, \pm 1, \ldots ; \quad n = 0, 1, \ldots, [T/\tau]-1$$

(voir fig. 8) et l'autre sur le réseau unidimensionnel

$$(x_m, 0) = (mh, 0), \quad m = 0, 1, \ldots$$

L'équation aux différences (4) est résoluble en u_m^{n+1} :

$$u_m^{n+1} = (1-r)u_m^n + ru_{m+1}^n + \tau\varphi(mh, n\tau). \tag{6}$$

Connaissant les valeurs u_m^n, $m = 0, \pm 1, \ldots$, de la solution $u^{(h)}$ aux points du réseau pour $t = n\tau$, on peut donc calculer les valeurs u_m^{n+1} pour $t = (n+1)\tau$. Les valeurs u_m^0 pour $t = 0$ étant données par les égalités $u_m^0 = \psi(mh)$, il est possible de calculer de proche en proche les valeurs de $u^{(h)}$ sur les droites $t = \tau$, $t = 2\tau$, ..., i.e. partout sur D_h.

Etablissons maintenant l'ordre d'approximation du schéma (5). On peut assimiler F_h à l'espace vectoriel de tous les couples de fonctions bornées $g^{(h)} = \begin{pmatrix}\varphi_m^n \\ \psi_m\end{pmatrix}$ en posant

$$\| g^{(h)} \|_{F_h} = \max_{m,n} | \varphi_m^n | + \max_m | \psi_m | \ (^*).$$

Comme indiqué au §13, la norme pour laquelle on considère l'approximation peut être introduite de façons différentes mais avec un arbitraire déterminé. Il nous suffit pour le moment de prendre comme norme la borne supérieure des modules de toutes les composantes de $g^{(h)} \in F_h$, et nous l'admettrons partout dans la suite de ce paragraphe.

Supposons la solution $u(x, t)$ du problème (4) possédant des dérivées secondes bornées. Selon la formule de Taylor on a alors

$$\left.\begin{aligned} \frac{u(x_m+h, t_n)-u(x_m, t_n)}{h} &= \frac{\partial u(x_m, t_n)}{\partial x} + \frac{h}{2}\,\frac{\partial^2 u(x_m+\xi, t_n)}{\partial x^2}, \\ \frac{u(x_m, t_n+\tau)-u(x_m, t_n)}{\tau} &= \frac{\partial u(x_m, t_n)}{\partial t} + \frac{\tau}{2}\,\frac{\partial^2 u(x_m, t_n+\eta)}{\partial t^2}, \end{aligned}\right\} \tag{7}$$

ξ et η étant des nombres dépendant de m, n et h et vérifiant les inégalités $0 < \xi < h$, $0 < \eta < \tau$.

Moyennant les formules (7) l'expression

$$L_h[u]_h \equiv \begin{cases} \dfrac{u(x_m, t_n+\tau)-u(x_m, t_n)}{\tau} - \dfrac{u(x_m+h, t_n)-u(x_m, t_n)}{h}, \\ u(x_m, 0) \end{cases}$$

(*) Si max $|\varphi_m^n|$ ou max $|\psi_m|$ n'est pas réalisé, il s'agit de la borne supérieure sup $|\varphi_m^n|$ ou sup $|\psi_m|$.

se récrit comme

$$L_h[u]_h = \begin{cases} \left(\dfrac{\partial u}{\partial t} - \dfrac{\partial u}{\partial x}\right)_{x_m,\, t_n} + \dfrac{\tau}{2}\,\dfrac{\partial^2 u(x_m,\, t_n+\eta)}{\partial t^2} - \dfrac{h}{2}\,\dfrac{\partial^2 u(x_m+\xi,\, t_n)}{\partial x^2}, \\ u(x_m,\, 0)+0 \end{cases}$$

ou

$$L_h[u]_h = f^{(h)} + \delta f^{(h)},$$

avec

$$\delta f^{(h)} = \begin{cases} \dfrac{\tau}{2}\,\dfrac{\partial^2 u(x_m,\, t_n+\eta)}{\partial t^2} - \dfrac{h}{2}\,\dfrac{\partial^2 u(x_m+\xi,\, t_n)}{\partial x^2}, \\ 0. \end{cases}$$

Donc

$$\| \delta f^{(h)} \|_{F_h} \leqslant \left(\sup \left| \frac{\partial^2 u}{\partial t^2} \right| \cdot \frac{r}{2} + \sup \left| \frac{\partial^2 u}{\partial x^2} \right| \cdot \frac{1}{2}\right) h.$$

Ainsi, le schéma aux différences considéré (5) approche le problème à l'ordre 1 en h pour la solution $u(x, t)$ admettant des dérivées secondes bornées.

3. Stabilité. Il nous reste à rappeler et à illustrer le concept de stabilité. Le problème aux limites aux différences (2) est stable par définition s'il existe des nombres $\delta > 0$ et $h_0 > 0$ tels que, quels que soient $h < h_0$ et $\delta f^{(h)}$ appartenant à F_h et vérifiant l'inégalité $\| \delta f^{(h)} \|_{F_h} < \delta$, le problème aux limites sous forme de différences

$$L_h z^{(h)} = f^{(h)} + \delta f^{(h)}$$

admette une solution et une seule et

$$\| z^{(h)} - u^{(h)} \|_{U_h} \leqslant C \, \| \delta f^{(h)} \|_{F_h},$$

C étant une constante indépendante de h.

En introduisant la notion de stabilité (§12) nous avons montré que, l'opérateur L_h étant linéaire, la définition formulée équivaut à dire :

Le problème aux limites aux différences (2) *est stable s'il existe un* $h_0 > 0$ *tel que, pour* $h < h_0$ *et* $f^{(h)} \in F_n$ *quelconque, il soit résoluble de façon unique et*

$$\| u^{(h)} \|_{U_h} \leqslant C \, \| f^{(h)} \|_{F_h}, \tag{8}$$

C étant une constante indépendante de h et $f^{(h)}$.

La propriété de stabilité s'interprète comme la sensibilité uniforme en h de la solution de (2) aux perturbations $\delta f^{(h)}$ du second membre.

Soulignons qu'en vertu de la définition citée, la stabilité est une propriété *intrinsèque* d'un problème aux différences, qui se formule indépendamment du problème différentiel, en particulier indépendamment de l'approximation et de la convergence.

Pourtant, *si un problème aux différences approche celui différentiel pour une solution u et est stable, il y a convergence* (3), *et l'ordre en h de la convergence coïncide avec l'ordre d'approximation.*

La démonstration de ce théorème important a été faite au §12.

Montrons la stabilité pour $r \leqslant 1$ du schéma aux différences (5). $\|\cdot\|_{U_h}$ sera muni d'une norme définie par l'égalité

$$\|u^{(h)}\|_{U_h} = \sup_{m,n} \|u_m^n\| = \max_n \sup_m |u_m^n|.$$

La norme $\|\cdot\|_{F_h}$ sera comprise au sens ci-dessus : si $g^{(h)} \in F_h$,

$$g^{(h)} = \begin{cases} \varphi_m^n, & m = 0, \pm 1, \ldots; \quad n = 0, 1, \ldots, [T/\tau], \\ \psi_m, & m = 0, \pm 1, \ldots, \end{cases}$$

alors

$$\|g^{(h)}\|_{F_h} = \max_{m,n} |\varphi_m^n| + \max_m |\psi_m| = \max_n \left[\max_m |\varphi_m^n| + \max_m |\psi_m|\right].$$

Mettons le problème aux différences

$$\left.\begin{aligned} \frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} &= \varphi_m^n, \quad m = 0, \pm 1, \ldots; \\ & \qquad n = 0, 1, \ldots, [T/\tau], \\ u_m^0 &= \psi_m, \quad m = 0, \pm 1, \ldots, \end{aligned}\right\} \tag{5'}$$

qui ne diffère de (5) que par le caractère arbitraire des seconds membres φ_m^n, ψ_m ne coïncidant en général pas avec $\varphi(mh, n\tau)$ et $\psi(mh)$, sous la forme

$$\left.\begin{aligned} u_m^{n+1} &= (1-r)u_m^n + ru_{m+1}^n + \tau\varphi_m^n, \\ u_m^0 &= \psi_m. \end{aligned}\right\} \tag{6'}$$

Puisque $r \leqslant 1$, on a $1 - r \geqslant 0$. Dans ce cas a lieu l'estimation

$$|(1-r)u_m^n + ru_{m+1}^n| \leqslant [(1-r)+r] \max(|u_m^n|, |u_{m+1}^n|) =$$
$$= \max(|u_m^n|, |u_{m+1}^n|) \leqslant \max_m |u_m^n|.$$

Utilisons-la et déduisons de (6')

$$|u_m^{n+1}| \leqslant \max_m |u_m^n| + \tau \max_m |\varphi_m^n| \leqslant \max_m |u_m^n| + \tau \max_{m,n} |\varphi_m^n|. \tag{6''}$$

Notons que si $\varphi_m^n \equiv 0$, l'inégalité (6'') entraîne la non-croissance de $\max_m |u_m^n|$ avec l'augmentation de n. Cette propriété d'un schéma aux différences porte le nom de *principe du maximum*. C'est ainsi que nous appellerons, afin d'être plus bref, l'inégalité tout entière

$$|u_m^{n+1}| \leqslant \max_m |u_m^n| + \tau \max_{m,n} |\varphi_m^n|.$$

Son second membre ne dépend pas de m de sorte qu'on peut effectuer dans le premier membre la substitution $|u_m^{n+1}| = \max_m |u_m^{n+1}|$ et obtenir

$$\max_m |u_m^{n+1}| \leqslant \max_m |u_m^n| + \tau \max_{m,n} |\varphi_m^n|.$$

De même

$$\max_m |u_m^n| \leqslant \max_m |u_m^{n-1}| + \tau \max_{m,n} |\varphi_m^n|,$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$\max_m |u_m^1| \leqslant \max_m |u_m^0| + \tau \max_{m,n} |\varphi_m^n|.$$

Une addition terme à terme et la réduction des termes semblables donnent

$$\max_m |u_m^{n+1}| \leqslant \max_m |u_m^0| + (n+1)\tau \max_{m,n} |\varphi_m^n|,$$

d'où immédiatement

$$\max_m |u_m^{n+1}| \leqslant \max_m |\psi_m| + T \max_{m,n} |\varphi_m^n| \leqslant \|f^{(h)}\|_{F_h} + T\|f^{(h)}\|_{F_h} = \\ = (1+T)\|f^{(h)}\|_{F_h}.$$

L'inégalité démontrée

$$\max_m |u_m^{n+1}| \leqslant (1+T)\|f^{(h)}\|_{F_h}$$

est valable pour tous les n de sorte qu'elle le reste en remplaçant $\max_m |u_m^{n+1}|$ par $\max_n \max_m |u_m^n| = \|u^{(h)}\|_{U_h}$:

$$\|u^{(h)}\|_{U_h} \leqslant (1+T)\|f^{(h)}\|_{F_h}. \tag{9}$$

L'inégalité (9) signifie la stabilité du problème linéaire (5) puisqu'il y a évidemment existence et unicité de la solution du problème (6′) pour φ_m^n et ψ_m arbitraires bornés. Le rôle de la constante C de l'inégalité (8) est ici échu à $1+T$.

Il serait faux de dire que l'approximation du problème aux limites différentiel (1) par le problème aux limites aux différences (2) assure à elle seule la stabilité et donc la convergence (3). Nous l'avons constaté au §9 sur l'exemple spécial d'un schéma aux différences approximant mais divergent.

Dans le cas d'équations aux dérivées partielles un schéma aux différences approximant pris au hasard s'avère en général inapplicable, ce qui fait que le choix d'un schéma stable (donc convergent) est le souci majeur du praticien du calcul numérique.

Rappelons par exemple que la démonstration de la stabilité du schéma aux différences (5) a été menée sous l'hypothèse $\tau/h = r \leqslant 1$. Si $r > 1$, le problème aux différences (5) approche toujours le problème (4) mais notre démonstration de la stabilité ne joue plus. Montrons que dans ce cas il n'y a pas de convergence de la solution $u^{(h)}$ du problème aux différences (5) vers la solution $u(x, t)$ du problème différentiel (4) et que la stabilité ne peut par conséquent avoir lieu, vu que la stabilité implique la convergence.

Soit, pour fixer les idées, $\varphi(x, t) \equiv 0$ de sorte que $\varphi(mh, n\tau)$ est également identiquement nul, et soit ensuite $T = 1$. Le pas h sera choisi de façon que le point (0, 1) du plan Oxt appartienne au réseau, i.e. le nombre

$$N = \frac{1}{\tau} = \frac{1}{rh}$$

est entier (fig. 9). En vertu de l'équation aux différences,

$$u_m^{n+1} = (1-r)u_m^n + ru_{m+1}^n.$$

La valeur $u_0^{n+1} = u_0^N$ de la solution $u^{(h)}$ au point (0, 1) du réseau s'exprime en fonction de ses valeurs u_0^n et u_1^n aux points $(0, 1-\tau)$ et $(h, 1-\tau)$, celles-ci en fonction des valeurs $u_0^{n-1}, u_1^{n-1}, u_2^{n-1}$ de la solution en trois points $(0, 1-2\tau)$, $(h, 1-2\tau)$ et $(2h, 1-2\tau)$. Les trois valeurs u_0^{n-1}, u_1^{n-1}, u_2^{n-1} s'expriment à leur tour par les valeurs de la solution en quatre points $(0, 1-3\tau)$, $(h, 1-3\tau)$, $(2h, 1-3\tau)$, $(3h, 1-3\tau)$, et ainsi de suite, si bien que la valeur finale u_0^{n+1} est représentée en fonction de u_m^0, valeurs de la solution aux points (0, 0), $(h, 0)$, $(2h, 0)$, ..., $(h/\tau, 0) = (Nh, 0)$ du réseau. Tous ces points se trouvent sur le segment

$$0 \leqslant x \leqslant \frac{h}{\tau} = \frac{1}{r}$$

de la droite $t = 0$ (fig. 9), où on a pour l'équation différentielle la condition initiale

$$u(x, 0) = \psi(x).$$

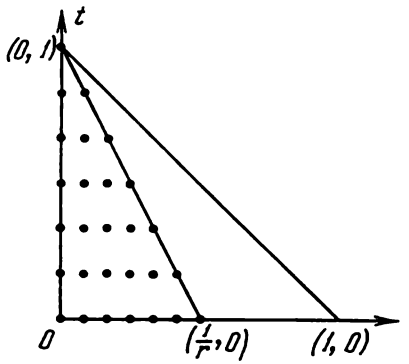


Fig. 9

Ainsi, la solution de l'équation aux différences au point (0, 1) du réseau ne dépend pas des valeurs de la fonction $\psi(x)$ aux points x extérieurs au segment

$$0 \leqslant x \leqslant \frac{1}{r}.$$

On vérifie aisément que le problème

$$\left.\begin{aligned} \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} &= 0, & -\infty < x < \infty, \quad t > 0, \\ u(x, 0) &= \psi(x), & -\infty < x < \infty, \end{aligned}\right\}$$

admet pour solution la fonction

$$u(x, t) \equiv \psi(x+t).$$

Celle-ci est constante sur chaque caractéristique $x+t = \text{const}$ et, en particulier, sur la droite $x+t = 1$ passant par les points (0, 1) et (1, 0) (voir fig. 9) et vaut $\psi(1)$ au point (0, 1). Il en résulte qu'il n'y a en général pas de convergence pour $r > 1$. En effet, le segment

$$0 \leqslant x \leqslant \frac{1}{r} < 1$$

de l'axe des abscisses ne renferme alors pas le point (1, 0). Si, pour une certaine $\psi(x)$, on avait fortuitement la convergence, on pourrait alors, sans

modifier $\psi(x)$ sur le segment

$$0 \leqslant x \leqslant \frac{1}{r}$$

et en conservant donc la valeur de la solution de l'équation aux différences au point (0, 1), violer la convergence en faisant varier $\psi(x)$ en $x = 1$ et au voisinage de ce point, fait qui se répercuterait sur la valeur $u(0, 1) = \psi(1)$ de la solution de l'équation différentielle. Cette variation de $\psi(x)$ pourrait être faite de façon qu'il y ait toujours existence de dérivées secondes de $\psi(x)$ et de la solution $u(x, t) = \psi(x+t)$ et donc approximation pour la solution $u(x, t)$. Dans ces conditions la stabilité du schéma (5) donnerait lieu à la convergence. Cependant, étant donné l'absence de convergence pour $r > 1$, la stabilité n'a pas lieu non plus.

La démonstration faite de l'instabilité du schéma aux différences (5) l'a été de façon indirecte. Il y a intérêt à suivre directement l'effet de l'instabilité du schéma (5) pour $r > 1$ sur la sensibilité de la solution $u^{(h)}$ aux erreurs à la donnée de $f^{(h)}$. Car c'est bien cette sensibilité uniforme par rapport à h qui a été définie par nous en tant que stabilité.

Admettons que, quel que soit h, on a les identités $\varphi(mh, n\tau) \equiv 0$ et $\psi(mh) \equiv 0$, de sorte que

$$f^{(h)} = \begin{Bmatrix} \varphi_m^n \\ \psi_m \end{Bmatrix} = 0,$$

et la solution $u^{(h)} = \{u_m^n\}$ du problème (5) est $u_m^n \equiv 0$. Supposons de plus qu'en se donnant les données initiales on a commis une erreur si bien qu'on s'est donné non pas $\psi_m = 0$ mais $\tilde{\psi}_m = (-1)^m \varepsilon$, $\varepsilon = \text{const}$, et donc, au lieu de

$$f^{(h)} = \begin{Bmatrix} \varphi_m^n \\ \psi_m \end{Bmatrix} = 0,$$

$$\tilde{f}^{(h)} = \begin{Bmatrix} 0 \\ \tilde{\psi}_m \end{Bmatrix}, \quad \|f^{(h)}\|_{F_h} = \varepsilon.$$

Désignons par $\tilde{u}^{(h)}$ la solution obtenue dans ces conditions. En vertu des équations

$$\tilde{u}_m^{n+1} = (1-r)\tilde{u}_m^n + r\tilde{u}_{m+1}^n,$$
$$\tilde{u}_m^0 = (-1)^m \varepsilon,$$

on a pour $\tilde{u}_m^1$:

$$\tilde{u}_m^1 = (1-r)\tilde{u}_m^0 + r\tilde{u}_{m+1}^0 = (1-r)(-1)^m\varepsilon + r(-1)^{m+1}\varepsilon =$$
$$= (1-2r)(-1)^m\varepsilon = (1-2r)\tilde{u}_m^0.$$

On voit que l'erreur commise pour $n = 0$ s'est trouvée amplifiée par le facteur $(1-2r)$. En passant à $\tilde{u}_m^2$ on obtient

$$\tilde{u}_m^2 = (1-r)\tilde{u}_m^1 + r\tilde{u}_{m+1}^1 = (1-2r)\tilde{u}_m^1 = (1-2r)^2\tilde{u}_m^0,$$

et en général

$$\tilde{u}_m^n = (1-2r)^n \tilde{u}_m^0 = (1-2r)^n (-1)^m \varepsilon.$$

Lorsque $r > 1$, on a $1-2r < -1$ de sorte qu'en passant d'un niveau $t = n\tau$ du réseau au niveau suivant l'erreur

$$\tilde{u}_m^0 = (-1)^m \varepsilon$$

est multipliée par un nombre négatif de module supérieur à l'unité. Pour $n = [T/\tau]$, on a

$$|\tilde{u}_m^n| = |1-2r|^{[T/\tau]} |\tilde{u}_m^0|.$$

D'où

$$||\tilde{u}^{(h)}||_{U_h} = |1-2r|^{[T/(rh)]} |\tilde{u}_m^0| = |1-2r|^{[T/(rh)]} \max |\tilde{\varphi}_m| =$$
$$= |1-2r|^{[T/(rh)]} ||\tilde{f}^{(h)}||_{F_h}.$$

Pour T fixe, l'erreur primitive $(-1)^m \varepsilon$ sur les données initiales augmente un nombre de fois égal à $|1-2r|^{[T/(rh)]}$, nombre qui croît rapidement avec h tendant vers zéro.

Faisons maintenant une brève critique de la façon dont nous avons apprécié la qualité de l'approximation en comparant la grandeur de la norme du résidu $||\delta f^{(h)}||_{F_h}$ avec une puissance de h. Il est connu que dans le cas de schémas stables l'ordre d'approximation coïncide avec l'ordre de grandeur de l'erreur $[u]_h - u^{(h)}$ sur la solution. La qualité d'un schéma s'évalue naturellement par le volume de calcul assurant la précision désirée, et le volume de calcul est en général proportionnel au nombre N de points du réseau utilisé. Pour les équations différentielles ordinaires N est proportionnel au pas h, si bien qu'en disant que l'erreur $\varepsilon \approx h^p$, nous affirmons par là même que $\varepsilon \approx 1/N^p$, c'est-à-dire qu'en réduisant l'erreur de moitié on augmente de $\sqrt[p]{2}$ fois le coût du calcul. Lorsqu'il s'agit des équations aux différences ordinaires, l'ordre d'approximation en h caractérise donc l'effort mathématique nécessaire.

Il n'en est plus de même en ce qui concerne les équations aux dérivées partielles. Dans l'exemple considéré d'un problème en deux variables x et t, le réseau est défini par deux pas τ et h. Le nombre N de points du réseau contenus dans un domaine borné du plan Oxt est en $1/(\tau h)$. Le nombre N peut lui aussi servir à évaluer le volume de calcul exigé par la résolution des équations aux différences. Soit $\tau = rh$. Alors $N \approx 1/h^2$, et $\varepsilon \approx h^p$ équivaut à $\varepsilon \approx 1/N^{p/2}$. Si $\tau = rh^2$, on a $N \approx 1/h^3$, et $\varepsilon \approx h^p$ équivaut à $\varepsilon \approx 1/N^{p/3}$.

Dans le cas des équations aux dérivées partielles il serait donc plus logique de mesurer l'ordre de grandeur de l'erreur non pas à l'aide des puissances de h mais par celles de $1/N$. Néanmoins nous fixons notre choix sur le procédé d'appréciation décrit vu qu'il est plus commode pour le calcul. Le lecteur ferait cependant bien de ne pas perdre de vue la circonstance notée lorsqu'il aura à apprécier la qualité des schémas aux différences.

Notons de plus que l'affirmation sur la proportionnalité de l'effort de calcul au nombre N de points du réseau n'est pas toujours vraie non plus.

On pourrait citer des exemples de schémas aux différences qui permettent d'obtenir une solution au bout de $\approx N^{1+q}$ opérations arithmétiques, avec $q = {}^1/_2$, 1 ou même 2. C'est le cas de problèmes aux limites aux différences approchant les équations elliptiques ou de problèmes en trois variables indépendantes ou plus (par exemple, $u = u(t, x, y)$). Pour le cas multidimensionnel la construction de schémas aux différences à l'aide desquels on obtient une solution par $\approx N$ opérations arithmétiques constitue un problème ardu (nous en parlerons aux §§31, 32).

S'il s'agit des calculs réels sur machine, on compare les algorithmes mis en œuvre en évaluant la qualité du schéma selon le critère temps machine. Celui-ci n'est pas forcément proportionnel au nombre d'opérations arithmétiques.

Il est à noter le rôle (parfois prédominant) du temps que demande le passage de l'information d'un bloc de mémoire à l'autre et le coût des opérations logiques.

Exercices

1. Pour le problème de Cauchy (4) étudier le schéma aux différences suivant :

$$\left.\begin{aligned} &\frac{u_m^{n+1}-u_m^n}{\tau}-\frac{u_m^n-u_{m-1}^n}{h}=\varphi(mh, n\tau),\\ &m=0, \pm 1, \ldots; \qquad n=0, 1, \ldots, [T/\tau]-1,\\ &u_m^0=\psi(mh), \qquad m=0, \pm 1, \ldots, \end{aligned}\right\}$$

où $\tau = rh$, $r = \text{const}$; à savoir

a) écrire l'opérateur L_h et le second membre $f^{(h)}$ pour ce schéma écrit comme $L_h u^{(h)} = f^{(h)}$,

b) donner la position relative de trois points du réseau en lesquels les valeurs de $u^{(h)}$ sont liées par l'équation aux différences pour m et n fixes,

c) montrer que le schéma aux différences approche le problème différentiel pour la solution $u(x, t)$ à l'ordre 1 en h, $u(x, t)$ possédant des dérivées secondes bornées,

d) vérifier si, pour un certain choix de r, $\tau = rh$, il y a stabilité du schéma étudié.

2. Pour le problème de Cauchy $u_t + u_x = \varphi(x, t)$, $u(x, 0) = \psi(x)$, $-\infty < x < \infty$, $0 \leqslant t \leqslant T$, étudier, suivant le plan proposé dans l'exercice précédent, chacun des schémas aux différences

$$\left.\begin{aligned} &\frac{u_m^{n+1}-u_m^n}{\tau}+\frac{u_m^n-u_{m-1}^n}{h}=\varphi(mh, n\tau),\\ &m=0, \pm 1, \ldots; \qquad n=0, 1, \ldots, [T/\tau]-1,\\ &u_m^0=\psi(mh), \qquad m=0, \pm 1, \ldots; \end{aligned}\right\}$$

$$\left.\begin{aligned} &\frac{u_m^{n+1}-u_m^n}{\tau}+\frac{u_{m+1}^n-u_m^n}{h}=\varphi(mh, n\tau),\\ &m=0, \pm 1, \ldots; \qquad n=0, 1, \ldots, [T/\tau]-1,\\ &u_m^0=\psi(mh), \qquad m=0, \pm 1, \ldots \end{aligned}\right\}$$

§22. Procédés simples de construction de schémas aux différences approximants

1. Remplacement des dérivées par des rapports aux différences. La technique la plus simple de construction de problèmes aux limites aux différences qui approchent les problèmes différentiels consiste à remplacer les dérivées par des rapports aux différences correspondants. On donne ci-dessous plusieurs exemples de schémas obtenus de cette façon où nous utiliserons les formules approchées

$$\left.\begin{aligned}
\frac{df(z)}{dz} &\approx \frac{f(z+\Delta z)-f(z)}{\Delta z},\\
\frac{df(z)}{dz} &\approx \frac{f(z)-f(z-\Delta z)}{\Delta z},\\
\frac{df(z)}{dz} &\approx \frac{f(z+\Delta z)-f(z-\Delta z)}{2\Delta z},\\
\frac{d^2f(z)}{dz^2} &\approx \frac{f(z+\Delta z)-2f(z)+f(z+\Delta z)}{\Delta z^2}.
\end{aligned}\right\} \tag{1}$$

Sous l'hypothèse que $f(z)$ ait un nombre suffisant de dérivées bornées, on peut écrire les expressions des restes de ces formules. Selon la formule de Taylor,

$$\left.\begin{aligned}
f(z+\Delta z) &= f(z)+\Delta z f'(z)+\frac{(\Delta z)^2}{2!}f''(z)+\frac{(\Delta z)^3}{3!}f'''(z)+\\
&\qquad +\frac{(\Delta z)^4}{4!}f^{(4)}(z)+o[(\Delta z)^4],\\
f(z-\Delta z) &= f(z)-\Delta z f'(z)+\frac{(\Delta z)^2}{2!}f''(z)-\frac{(\Delta z)^3}{3!}f'''(z)+\\
&\qquad +\frac{(\Delta z)^4}{4!}f^{(4)}(z)+o[(\Delta z)^4].
\end{aligned}\right\} \tag{2}$$

Les développements (2) permettent de représenter les restes de (1), notamment on a

$$\left.\begin{aligned}
\frac{f(z+\Delta z)-f(z)}{\Delta z} &= f'(z)+\left[\frac{\Delta z}{2}f''(z)+o(\Delta z)\right],\\
\frac{f(z)-f(z-\Delta z)}{\Delta z} &= f'(z)+\left[-\frac{\Delta z}{2}f''(z)+o(\Delta z)\right],\\
\frac{f(z+\Delta z)-f(z-\Delta z)}{2\Delta z} &= f'(z)+\left[\frac{(\Delta z)^2}{3}f'''(z)+o((\Delta z)^2)\right],\\
\frac{f(z+\Delta z)-2f(z)+f(z-\Delta z)}{\Delta z^2} &= f''(z)+\left[\frac{(\Delta z)^2}{12}f^{(4)}(z)+o((\Delta z)^2)\right].
\end{aligned}\right\} \tag{3}$$

Les restes des formules approchées (1) figurent dans les égalités correspondantes (3) sous forme d'expressions entre crochets.

Il est évident que les formules (1) et les expressions (3) des restes peuvent être utilisées en remplaçant les dérivées partielles par des rapports

aux différences. Par exemple,

$$\frac{\partial u(x, t)}{\partial t} \approx \frac{u(x, t+\Delta t)-u(x, t)}{\Delta t}$$

et

$$\frac{u(x, t+\Delta t)-u(x, t)}{\Delta t} = \frac{\partial u(x, t)}{\partial t} + \left[\frac{\Delta t}{2}\,\frac{\partial^2 u(x, t)}{\partial t^2} + o(\Delta t)\right].$$

De même

$$\frac{\partial u(x, t)}{\partial x} \approx \frac{u(x+\Delta x, t)-u(x, t)}{\Delta x}$$

et

$$\frac{u(x+\Delta x, t)-u(x, t)}{\Delta x} = \frac{\partial u(x, t)}{\partial x} + \left[\frac{\Delta x}{2}\,\frac{\partial^2 u(x, t)}{\partial x^2} + o(\Delta x)\right],$$

et ainsi de suite.

Exemple 1. Reprenons le problème de Cauchy (4) du §21 :

$$\left.\begin{aligned} \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} &= \varphi(x, t), \qquad -\infty < x < \infty, \quad 0 \leqslant t \leqslant T, \\ u(x, 0) &= \psi(x), \qquad -\infty < x < \infty. \end{aligned}\right\} \tag{4}$$

Construisons pour l'approcher trois schémas en utilisant dans les trois cas le réseau D_h formé par les points d'intersection des droites $x = mh$, $t = n\tau$ situés dans la bande $0 \leqslant t \leqslant T$. Les valeurs de τ et h seront supposées reliées par l'égalité $\tau = rh$, avec r une constante positive. Le plus simple de ces schémas s'écrit comme (5) du §21

$$L_h u^{(h)} \equiv \left\{\begin{aligned} &\frac{u_m^{n+1}-u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = \varphi(mh, n\tau), \\ &\qquad\qquad u_m^0 = \psi(mh), \end{aligned}\right. \tag{5}$$

et s'obtient en remplaçant les dérivées $u_t = \partial u/\partial t$ et $u_x = \partial u/\partial x$ par les formules approchées

$$u_t(x, t) \approx \frac{u(x, t+\tau)-u(x, t)}{\tau},$$

$$u_x(x, t) \approx \frac{u(x+h, t)-u(x, t)}{h}.$$

Une étude détaillée de ce schéma a été faite au §21. Dans son cas le résidu $\delta f^{(h)}$ apparu en substituant la solution $[u]_h$ du problème différentiel au premier membre du problème aux différences

$$L_h[u]_h = f^{(h)} + \delta f^{(h)}$$

est défini par la formule

$$\delta f^{(h)} = \begin{cases} \left(\dfrac{\tau}{2}\,u_{tt} - \dfrac{h}{2}\,u_{xx}\right)_m^n + o(\tau+h), \\ 0. \end{cases}$$

La norme d'un élément $f^{(h)} \in F_h$ sera assimilée dans ce paragraphe au maximum de toutes les composantes de $f^{(h)}$. On a alors évidemment

$$\|\delta f^{(h)}\|_{F_h} = O(\tau+h) = O(rh+h) = O(h),$$

et l'approximation est d'ordre 1.

Le deuxième schéma s'obtient en recourant à une autre formule pour $\partial u/\partial x$:

$$\frac{\partial u(x, t)}{\partial x} \approx \frac{u(x, t)-u(x-h, t)}{h},$$

et il est de la forme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1}-u_m^n}{\tau} - \dfrac{u_m^n - u_{m-1}^n}{h} = \varphi(mh, n\tau), \\ u_m^0 = \psi(mh). \end{cases}$$

Ici

$$\delta f^{(h)} \equiv \begin{cases} \left(\dfrac{\tau}{2} u_{tt} + \dfrac{h}{2} u_{xx}\right)_m^n + o(\tau+h), \\ 0, \end{cases}$$

$$\|\delta f^{(h)}\|_{F_h} = O(h),$$

et l'approximation est de nouveau d'ordre 1.

La différence entre ces schémas paraît inessentielle. Mais nous verrons dans la suite que le deuxième schéma est impraticable : il est instable pour tout $\tau/h = r = \text{const}$.

On aboutit au troisième schéma

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{n+1} - \dfrac{u_{m+1}^n + u_{m-1}^n}{2}}{\tau} - \dfrac{u_{m+1}^n - u_m^n}{h} = \varphi(mh, n\tau), \\ u_m^0 = \psi(mh), \end{cases}$$

en remplaçant les dérivées par des rapports aux différences moyennant les formules approchées

$$\frac{\partial u(x, t)}{\partial t} \approx \frac{u(x, t+\tau) - \dfrac{u(x+h, t)+u(x-h, t)}{2}}{\tau},$$

$$\frac{\partial u(x, t)}{\partial x} \approx \frac{u(x+h, t)-u(x, t)}{h}.$$

Si la solution $u(x, t)$ du problème (1) est suffisamment régulière, on trouve à l'aide des développements tayloriens (2)

$$\frac{u(x, t+\tau) - \dfrac{u(x+h, t)+u(x-h, t)}{2}}{\tau} - \frac{u(x+h, t)-u(x, t)}{h} =$$

$$= \left[\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} - \frac{h^2}{2\tau}\frac{\partial^2 u}{\partial x^2} + \frac{\tau}{2}\frac{\partial^2 u}{\partial t^2}\right]_{x,t} + O\left(\tau^2 + h^2 + \frac{h^4}{\tau}\right) =$$

$$= \varphi(x, t) + \left[-\frac{h}{2r} u_{xx} + \frac{\tau}{2} u_{tt} + O(h^2)\right]_{x,t}.$$

Donc

$$L_h[u]_h = \begin{cases} \varphi(mh, nh) + \left[-\frac{h}{2r} u_{xx} + \frac{\tau}{2} u_{tt} + O(h^2)\right], \\ \psi(mh) + 0, \end{cases}$$

de sorte que $\delta f^{(h)}$ de la formule

$$L_h[u]_h = f^{(h)} + \delta f^{(h)}$$

est de la forme

$$\delta f^{(h)} \equiv \begin{cases} -\frac{h}{2r} u_{xx} + \frac{\tau}{2} u_{tt} + O(h^2), \\ 0. \end{cases}$$

Par conséquent, $\|\delta f^{(h)}\|_{F_h} = O(h)$ et l'ordre d'approximation est 1 comme dans les deux exemples précédents.

Plaçons-nous maintenant dans le cas où il y a entre les pas du réseau une relation définie non pas par $\tau = rh$, mais par

$$\tau = rh^2, \; r = \text{const},$$

déterminant une diminution accélérée du pas τ devant le pas h. Alors

$$L_h[u]_h = \begin{cases} \left[\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} - \frac{1}{2r} \frac{\partial^2 u}{\partial x^2}\right]_{x_m, t_n} + O(h^2), \\ u(mh, n\tau). \end{cases}$$

On en conclut que le schéma aux différences considéré approche le problème

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} - \frac{1}{2r} \frac{\partial^2 u}{\partial x^2} = \varphi(x, t),$$

$$u(x, 0) = \psi(x)$$

et non pas le problème de Cauchy (4) dont nous nous sommes proposé d'obtenir une approximation.

Ainsi, un même schéma aux différences approche pour $h \to 0$ des problèmes différentiels différents selon la relation $\tau = \tau(h)$. De tels schémas sont dits *rigides*.

On facilite le stockage d'un schéma aux différences en le représentant sous forme d'une image donnant la position relative des points du réseau (pochoir) en lesquels les valeurs de la solution sont liées par l'équation aux différences pour certains m et n fixes.

La figure 10 donne les pochoirs correspondant aux trois schémas étudiés.

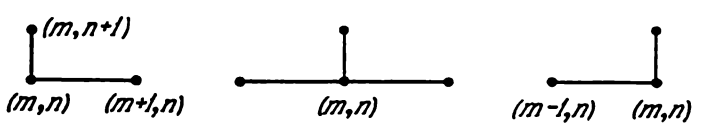


Fig. 10

EXEMPLE 2. Considérons deux schémas aux différences approchant le problème de Cauchy pour l'équation de la chaleur

$$\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}=\varphi(x, t), \qquad -\infty < x < \infty, \quad 0 < t < T,$$
$$u(x, 0) = \psi(x), \qquad -\infty < x < \infty.$$

Le schéma le plus simple

$$L_h^{(1)}u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1}-u_m^n}{\tau}-\dfrac{u_{m+1}^n-2u_m^n+u_{m-1}^n}{h^2} = \varphi(mh, n\tau), \\ u_m^0 = \psi(mh), \end{cases}$$
$$f^{(h)} \equiv \begin{cases} \varphi(mh, n\tau), \\ \psi(mh) \end{cases}$$

s'obtient en remplaçant les dérivées u_t et u_{xx} par des rapports aux différences conformément aux formules

$$u_t(x, t) \approx \frac{u(x, t+\tau)-u(x, t)}{\tau},$$
$$u_{xx}(x, t) \approx \frac{u(x+h, t)-2u(x, t)+u(x-h, t)}{h^2}.$$

A condition d'utiliser pour $u_{xx}(x, t)$ une autre formule

$$u_{xx}(x, t) \approx \frac{u(x+h, t+\tau)-2u(x, t+\tau)+u(x-h, t+\tau)}{h^2},$$

nous arrivons à un autre schéma

$$L_h^{(2)}u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1}-u_m^n}{\tau}-\dfrac{u_{m+1}^{n+1}-2u_m^{n+1}+u_{m-1}^{n+1}}{h^2} = \varphi(mh, n\tau), \\ u_m^0 = \psi(mh) \end{cases}$$

pour la même équation. Afin de pouvoir distinguer les opérateurs L_h correspondants, nous les avons numérotés : $L_h^{(1)}u^{(h)} = f^{(h)}$, $L_h^{(2)}u^{(h)} = f^{(h)}$. Les pochoirs associés à ces schémas sont représentés à la figure 11.

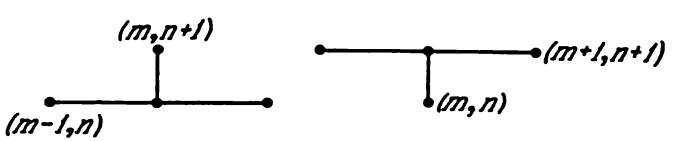


Fig. 11

Ces schémas diffèrent sensiblement. En effet, en calculant la solution $u^{(h)}$ par le premier on ne se heurte à aucune difficulté et on utilise la formule explicite

$$u_m^{n+1} = (1-2r)u_m^n+r(u_m^n+u_{m+1}^n)+\tau\varphi(mh, n\tau),$$

où $r = \tau/h^2$. Celle-ci s'obtient de l'équation aux différences qu'on résout en u_m^{n+1}. Connaissant les valeurs u_m^n, $m = 0, \pm 1, \ldots$, de la solution au

niveau $t = t_n (= n\tau)$ du réseau, on peut trouver ses valeurs u_m^{n+1} au niveau suivant $t = t_{n+1}$.

Cette propriété commode n'est plus celle du deuxième schéma et l'équation aux différences écrite pour m et n fixes n'est pas résoluble en u_m^{n+1} en exprimant cette valeur par les valeurs connues u_{m+1}^n, u_m^n, u_{m-1}^n du niveau précédent. Le fait est que cette équation contient, en plus de la valeur inconnue u_m^{n+1}, les inconnues u_{m-1}^{n+1}, u_{m+1}^{n+1}. On définit donc u_m^{n+1}, $m = 0, \pm 1, \ldots$, en résolvant l'équation aux différences relativement à la fonction u_m^{n+1} de l'argument m. Malgré tout ce que nous venons de dire, nous montrerons dans la suite que le schéma $L_h^{(2)}u^{(h)} = f^{(h)}$ est, comme de règle, plus commode que $L_h^{(1)}u^{(h)} = f^{(h)}$.

Lorsque $\tau = rh^2$, $r = \text{const}$, les deux schémas réalisent une approximation d'ordre deux en h. Calculons le résidu $\delta f^{(h)}$ et évaluons l'ordre d'approximation du deuxième schéma. Moyennant les formules (3) nous pouvons écrire

$$L_h^{(2)}[u]_h \equiv \begin{cases} (u_t - u_{xx})_{\substack{x=mh,\\ t=(n+1)\tau}} - \dfrac{\tau}{2} u_{tt}(x, t_{n+1}) - \dfrac{h^2}{12} u_{xxxx}(x, t_{n+1}) + o(\tau + h^2), \\ u(mh, 0). \end{cases}$$

D'où, compte tenu de $\tau = rh^2$,

$$L_h^{(2)}[u]_h \equiv \begin{cases} \varphi(x_m, t_{n+1}) + O(h^2), \\ \psi(mh) + 0, \end{cases}$$

$$\delta f^{(h)} = \begin{cases} \varphi(x_m, t_n) - \varphi(x_m, t_{n+1}) + O(h^2), \\ 0. \end{cases}$$

Mais

$$\varphi(x_m, t_{n+1}) = \varphi(x_m, t_n) + [\varphi(x_m, t_{n+1}) - \varphi(x_m, t_n)] = \\ = \varphi(x_m, t_n) + O(\tau) = \varphi(x_m, t_n) + O(h^2).$$

C'est pourquoi

$$\|\delta f^{(h)}\|_{F_h} = O(h^2).$$

Exemple 3. Considérons le plus simple schéma aux différences qui approche le problème de Dirichlet relatif à l'équation de Poisson dans le carré $D(0 < x < 1, 0 < y < 1)$ de frontière Γ (fig. 12, *a*) :

$$u_{xx} + u_{yy} = \varphi(x, y), \quad (x, y) \in D,$$
$$u|_\Gamma = \psi(x, y), \quad (x, y) \in \Gamma.$$

Construisons un réseau D_h avec les points $(x_m, t_n) = (mh, nh)$ de l'intérieur ou du contour du carré. Le pas h sera choisi tel que le nombre $1/h$ soit entier. Définissons le schéma aux différences $L_h u^{(h)} = f^{(h)}$ par les

égalités

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{m+1,n}-2u_{mn}+u_{m-1,n}}{h^2}+\dfrac{u_{m,n+1}-2u_{mn}+u_{m,n-1}}{h^2} = \\ \qquad = \varphi(mh, nh), \quad (mh, nh) \in D, \\ u_{mn} = \psi(mh, nh), \quad (mh, nh) \in \Gamma, \end{cases}$$

$$f^{(h)} = \begin{cases} \varphi(mh, nh) \text{ si } (mh, nh) \in D, \\ \psi(mh, nh) \text{ si } (mh, nh) \in \Gamma. \end{cases}$$

Le résidu $\delta f^{(h)}$, $L_h[u]_h = f^{(h)}+\delta f^{(h)}$, s'écrit, en vertu des formules (3),

$$\delta f^{(h)} = \begin{cases} \dfrac{h^2}{12}(u_{xxxx}+u_{yyyy})|_{x_m, y_n}+o(h^2), \\ 0, \end{cases}$$

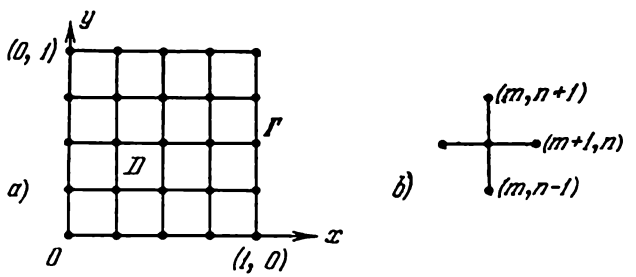


Fig. 12

de sorte qu'on a une approximation d'ordre 2. Le pochoir à cinq points associé à l'équation aux différences utilisée est représenté à la figure 12, *b*.

Les schémas aux différences ci-dessus s'obtenaient en remplaçant chaque dérivée de l'équation différentielle par un rapport aux différences.

2. Méthode des coefficients indéterminés. Un procédé plus général de formation de schémas aux différences consiste à approcher tout l'opérateur différentiel et non pas chaque dérivée prise à part. Voyons cette méthode pour plusieurs schémas aux différences pour le problème de Cauchy (4) et prenons en premier lieu le schéma (5) d'ordre d'approximation un. Ce schéma relie les valeurs de la fonction cherchée aux trois points représentés à la figure 10 à gauche. L'équation aux différences utilisée

$$\Lambda_h u^{(h)} \equiv \frac{u_m^{n+1}-u_m^n}{\tau}-\frac{u_{m+1}^n-u_m^n}{h} = \varphi(mh, n\tau)$$

est de la forme

$$\Lambda_h u^{(h)} \equiv a^0 u_m^{n+1}+a_0 u_m^n+a_1 u_{m+1}^n = \varphi(mh, n\tau).$$

Oublions momentanément que nous connaissons déjà le schéma aux différences (5) pour lequel

$$a^0 = \frac{1}{\tau}, \quad a_0 = \frac{1}{h}-\frac{1}{\tau}, \quad a_1 = -\frac{1}{h}$$

et, estimant ces coefficients indéterminés, essayons de les choisir de façon qu'on ait

$$\Lambda_h[u]_h\Big|_{\substack{x=mh,\\ t=n\tau}} = \left(\frac{\partial u}{\partial t}-\frac{\partial u}{\partial x}\right)\Big|_{\substack{x=mh,\\ t=n\tau}}+O(h)$$

ou

$$\Lambda_h[u]_h\Big|_{\substack{x=mh,\\ t=n\tau}} = \Lambda u\Big|_{\substack{x=mh,\\ t=n\tau}}+O(h) \tag{6}$$

avec

$$\Lambda u \equiv \frac{\partial u}{\partial t}-\frac{\partial u}{\partial x}. \tag{7}$$

Utilisons la formule de Taylor :

$$u[mh, (n+1)\tau] = u(mh, n\tau)+\tau u'_t(mh, n\tau)+O(\tau^2),$$
$$u[(m+1)\,h, n\tau] = u(mh, n\tau)+hu'_x(mh, n\tau)+O(h^2).$$

Portons ces expressions au second membre de l'égalité

$$\Lambda_h[u]_h\Big|_{\substack{x=mh,\\ t=n\tau}} \equiv a^0u[mh, (n+1)\tau]+a_0u(mh, n\tau)+a_1u[(m+1)h, n\tau],$$

il vient

$$\Lambda_h[u]_h\Big|_{\substack{x=mh,\\ t=n\tau}} = (a^0+a_0+a_1)\,u(mh, n\tau)+a^0\tau\,\frac{\partial u(mh, n\tau)}{\partial t}+$$
$$+a_1h\,\frac{\partial u(mh, n\tau)}{\partial x}+O(a^0\tau^2, a_1h^2). \tag{8}$$

Puisqu'on s'est proposé de choisir les coefficients a^0, a_0, a_1 de façon que soit remplie la condition d'approximation (6), il est naturel de grouper préalablement les termes du second membre de (8) pour mettre en évidence le terme (7). Les termes restants constituent alors un reste qui doit être petit. On isole le terme Λu en remplaçant les dérivées $\partial u/\partial t$ et $\partial u/\partial x$ du second membre de (8) par les formules respectives

$$\frac{\partial u}{\partial t} \equiv \Lambda u+\frac{\partial u}{\partial x} \quad \text{et} \quad \frac{\partial u}{\partial x} = \frac{\partial u}{\partial t}-\Lambda u.$$

Pour fixer les idées, arrêtons-nous sur la première.

Les pas τ et h seront de plus reliés par $\tau = rh$, r étant une constante. Dans ces conditions, l'égalité (8) prend la forme

$$\Lambda_h[u]_h\Big|_{\substack{x=mh,\\ t=n\tau}} = a^0rh\Lambda u\Big|_{\substack{x=mh,\\ t=n\tau}}+(a^0+a_0+a_1)\,u(mh, n\tau)+$$
$$+(a^0r+a_1)\,hu_x(mh, n\tau)+O(a^0r^2h^2, a_1h^2). \tag{9}$$

Parmi toutes les fonctions régulières $u(x, t)$ il en existe pour lesquelles u, $\partial u/\partial x$ et $\partial u/\partial t$ prennent en tout point fixe donné à l'avance des valeurs indépendantes quelconques. On peut donc affirmer l'indépendance des valeurs

$$u, \frac{\partial u}{\partial x} \quad \text{et} \quad \Lambda u = \frac{\partial u}{\partial t}-\frac{\partial u}{\partial x} = \varphi(x, t).$$

Il résulte, par conséquent, de l'égalité (9) que pour que tout second membre $\varphi(x, t)$ du problème (4) vérifie la condition d'approximation

$$\Lambda_h[u]_h\,|_{\substack{x=mh,\\ t=n\tau}} = (\Lambda u)_{\substack{x=mh,\\ t=n\tau}} + O(h)$$

il faut qu'on ait

$$\begin{aligned} a^0rh &= 1+O_1(h),\\ a^0+a_0+a_1 &= 0+O_2(h),\\ (a^0r+a_1)h &= 0+O_3(h), \end{aligned}$$

avec $O_1(h)$, $O_2(h)$, $O_3(h)$ des quantités en h. Posons $O_1(h) = O_2(h) = O_3(h) = 0$. Le système qui s'obtient alors

$$\begin{aligned} a^0rh &= 1,\\ a^0+a_0+a_1 &= 0,\\ a^0r+a_1 &= 0 \end{aligned}$$

admet une solution unique

$$\begin{aligned} a^0 &= \frac{1}{rh} = \frac{1}{\tau},\\ a_0 &= \frac{r-1}{rh} = \frac{1}{h}-\frac{1}{\tau},\\ a_1 &= -\frac{1}{h}, \end{aligned}$$

qui conduit au schéma connu (5).

Nous savons maintenant que c'est l'unique schéma parmi des schémas de la forme

$$L_hu^{(h)} \equiv \begin{cases} a^0u_m^{n+1}+a_0u_m^n+a_1u_{m+1}^n = \varphi(mh, n\tau),\\ u_m^0 = \psi(mh) \end{cases}$$

qui approche le problème de Cauchy considéré. Disant « unique » nous négligeons l'arbitraire dû au choix libre des fonctions $O_1(h)$, $O_2(h)$, $O_3(h)$. Nous agirons de même dans tous les exemples qui suivront et nous poserons parfois, dès le début, égales à zéro des quantités arbitraires analogues à $O_1(h)$, $O_2(h)$, $O_3(h)$.

Le lecteur vérifiera sans peine qu'une fois prises en considération ces quantités conduiraient à une modification peu importante sur le résultat de l'exemple considéré :

$$\begin{aligned} a^0 &= \frac{1}{h}\left[\frac{1}{r}+O(h)\right],\\ a_0 &= \frac{1}{h}\left[\frac{r-1}{r}+O(h)\right],\\ a_1 &= \frac{1}{h}[-1+O(h)]. \end{aligned}$$

Il en sera de même pour d'autres exemples que nous aurons à étudier. Construisons maintenant pour (4) les schémas aux différences

$$L_h u^{(h)} \equiv \begin{cases} a^0 u_m^{n+1} + a_0 u_m^n + a_{-1} u_{m+1}^n + a_1 u_{m+1}^n = \varphi(mh, n\tau), \\ u_m^0 = \psi(mh) \end{cases} \tag{10}$$

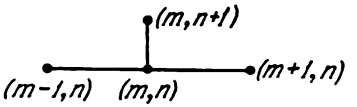


Fig. 13

plus généraux reliant les valeurs de la fonction cherchée aux quatre points de la figure 13.

Imposons toujours aux pas du réseau la condition $\tau = rh$, $r = \text{const}$, et introduisons la notation Λ_h en posant

$$\Lambda_h u^{(h)} \equiv a^0 u_m^{n+1} + a_0 u_m^n + a_{-1} u_{m-1}^n + a_1 u_{m+1}^n. \tag{11}$$

Pour toute fonction $u(x, t)$ suffisamment régulière on écrit moyennant la formule de Taylor

$$\begin{aligned} \Lambda_h[u]_h|_{\substack{x=mh,\\ t=n\tau}} = {} & (a^0 + a_0 + a_1 + a_{-1})\, u(mh, n\tau) + \\ & + a^0 rh u_t(mh, n\tau) + (a_1 - a_{-1})\, h u_x(mh, n\tau) + \\ & + \frac{1}{2} a^0 r^2 h^2 u_{tt}(mh, n\tau) + \frac{1}{2}(a_1 + a_{-1})\, h^2 u_{xx}(mh, n\tau) + \\ & + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3). \end{aligned} \tag{12}$$

Mettons en évidence dans le second membre le terme $\Lambda u \equiv (\partial u/\partial t) - (\partial u/\partial x)$ en utilisant l'identité $u_t = u_x + \Lambda u$. Il vient

$$\begin{aligned} \Lambda_h[u]_h|_{\substack{x=mh,\\ t=n\tau}} = {} & a^0 rh \Lambda u|_{\substack{x=mh,\\ t=n\tau}} + (a^0 + a_0 + a_1 + \\ & + a_{-1})\, u(mh, n\tau) + (a^0 r + a_1 - a_{-1})\, h u_x(mh, n\tau) + \\ & + \frac{1}{2} a^0 r^2 h^2 u_{tt}(mh, n\tau) + \frac{1}{2}(a_1 + a_{-1})\, h^2 u_{xx}(mh, n\tau) + \\ & + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3). \end{aligned}$$

Sous l'hypothèse de $O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)$ suffisamment petit — on verra dans la suite qu'il en est bien ainsi, — pour que la condition d'approximation

$$(\Lambda_h[u]_h)_{\substack{x=mh,\\ t=n\tau}} = (\Lambda u)_{\substack{x=mh,\\ t=n\tau}} + O(h)$$

ait lieu, il faut que les nombres a^0, a_0, a_1, a_{-1} vérifient les trois égalités suivantes :

$$\begin{aligned} a^0 rh &= 1 + O_1(h), \\ a^0 + a_0 + a_1 + a_{-1} &= 0 + O_2(h), \\ (a^0 r + a_1 - a_{-1}) h &= 0 + O_3(h). \end{aligned}$$

Comme nous sommes convenus, posons $O_1(h) = 0$, $O_2(h) = 0$, $O_3(h) = 0$. Il vient le système

$$\left.\begin{aligned} a^0rh &= 1, \\ a^0+a_0+a_1+a_{-1} &= 0, \\ a^0r+a_1-a_{-1} &= 0. \end{aligned}\right\} \tag{13}$$

Si ces conditions sont remplies, alors

$$\Lambda_h[u]_h|_{\substack{x=mh,\\ t=n\tau}} = \Lambda u|_{\substack{x=mh,\\ t=n\tau}} + \frac{1}{2}a^0r^2h^2u_{tt}(mh, n\tau)+$$
$$+\frac{1}{2}(a_1+a_{-1})h^2u_{xx}(mh, n\tau)+O(a^0r^3h^3, a_1h^3, a_{-1}h^3).$$

Le système (13) admet une famille de solutions dépendant d'un paramètre. L'une de ces solutions

$$a^0 = \frac{1}{rh}, \quad a_{-1} = 0, \quad a_0 = \frac{r-1}{h}, \quad a_1 = -\frac{1}{h}$$

donne le schéma considéré (5). A la solution

$$a^0 = \frac{1}{rh}, \quad a_0 = -\frac{1}{rh}, \quad a_{-1} = \frac{1}{2h}, \quad a_1 = -\frac{1}{2h}$$

correspond le schéma

$$L_hu^{(h)} = \begin{cases} \dfrac{u_m^{(n+1)}-u_m^n}{\tau} - \dfrac{u_{m+1}^n-u_{m-1}^n}{2h} = \varphi(mh, n\tau), \\ u_m^0 = \psi(mh). \end{cases}$$

Ayant choisi une solution du système (13) il faut la porter dans le reste et s'assurer que celui-ci est petit. S'agissant des deux solutions considérées la substitution de a^0, a_0, a_1, a_{-1} donne les restes

$$\frac{a^0r^2h^2}{2}\frac{\partial^2u}{\partial t^2}+\frac{a_1+a_{-1}}{2}h^2\frac{\partial^2u}{\partial x^2}+O(a^0r^3h^3, a_1h^3, a_{-1}h^3)$$

en $O(h)$.

Parmi $u(x, t)$ régulières il existe des polynômes du deuxième degré tels que $\partial^2u/\partial t^2$ et $\partial^2u/\partial x^2$ prennent en tout point fixe n'importe quelles valeurs indépendantes données à l'avance. Cela étant, le terme $O(a^0r^3h^3, a_1h^3, a_{-1}h^3)$ contenant les dérivées troisièmes du polynôme $u(x, t)$ devient nul. Ainsi, pour que le reste soit en h, il faut que chacun des coefficients de $\partial^2u/\partial t^2$ et $\partial^2u/\partial x^2$ le soit également. Puisque la première équation (13) donne $a^0 = 1/(rh)$, le coefficient de $\partial^2u/\partial t^2$ est $rh/2$ et le reste est toujours de l'ordre de 1 au plus.

Nous avons établi l'impossibilité de construire un schéma aux différences de la forme (10) qui approche le problème

$$Lu \equiv \begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial u}{\partial x} = \varphi(x, t), \\ u(x, 0) = \psi(x) \end{cases}$$

à l'ordre h^2. Une élévation de l'ordre d'approximation exigerait un nombre plus grand de points du réseau aux différences par rapport à celui qu'utilise le schéma construit.

Il existe cependant un procédé de construction d'un schéma donnant une approximation de l'ordre de h^2 et se contentant des quatre points indiqués du réseau. Ce procédé que nous allons illustrer sur un exemple porte un caractère général. Il se trouve qu'on peut prendre les coefficients a^0, a_{-1}, a_0, a_1 tels qu'on ait l'égalité

$$\begin{aligned}\Lambda_h[u]_h \equiv\ & a^0u(mh, (n+1)\tau)+a_{-1}u((m-1)h, n\tau)+\\ & +a_0u(mh, n\tau)+a_1u((m+1)h, n\tau) =\\ & = \Lambda u+\frac{rh}{2}[(\Lambda u)_t+(\Lambda u)_x]\Big|_{\substack{x=mh,\\ t=n\tau}}+\\ & +O(h^2) = P_h\Lambda u\big|_{(x_m,\, t_n)}+O(h^2),\end{aligned}$$

où

$$P_h = E+\frac{rh}{2}\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}\right),$$

E étant l'opérateur de multiplication par l'unité. Vu que $\Lambda u = u_t-u_x = \varphi(x, t)$, le schéma aux différences

$$L_hu^{(h)} = \begin{cases} a^0u_m^{n+1}+a_{-1}u_{m-1}^n+a_0u_m^n+a_1u_{m+1}^n = \varphi_m^n, \\ u_m^0 = \psi(mh),\end{cases}$$

avec

$$\varphi_m^n = (P_h\varphi)_{\substack{x=mh,\\ t=n\tau}} = \varphi+\frac{rh}{2}(\varphi_t+\varphi_x)\Big|_{\substack{x=mh,\\ t=n\tau}},$$

approchera alors le problème différentiel considéré pour la solution $u(x, t)$ à l'ordre deux en h.

Les coefficients a^0, a_{-1}, a_0, a_1 peuvent être choisis, cette fois encore, par la méthode des coefficients indéterminés. Ils sont alors

$$a^0 = \frac{1}{rh},\quad a_0 = -\frac{1}{rh}+\frac{r}{h},\quad a_{-1} = \frac{1-r}{2h},\quad a_1 = -\frac{1+r}{2h}.$$

et l'opérateur Λ_h s'écrit

$$\Lambda_hu^{(h)} = \frac{1}{\tau}(u_m^{n+1}-u_m^n)-\frac{1}{2h}(u_{m+1}^n-u_{m-1}^n)-\frac{r}{2h}(u_{m+1}^n-2u_m^n+u_{m-1}^n).$$

La méthode des coefficients indéterminés permet non seulement de choisir a^0, a_{-1}, a_0, a_1, tels que

$$\begin{aligned}\Lambda_h[u]_h = \ & a^0u(x, t+\tau)+a_{-1}u(x-h, t)+a_0u(x, t)+\\ & +a_1u(x+h, t) = P_h\Lambda u+O(h^2)\end{aligned}$$

pour l'opérateur P_h écrit plus haut, mais aussi de construire tous ces opérateurs.

Voici comment on procède.
Admettant que

$$\Lambda_h u^{(h)} \equiv a^0 u_m^{n+1} + a_{-1} u_{m-1}^n + a_0 u_m^n + a_1 u_{m+1}^n,$$

on obtient, à l'aide de la formule de Taylor,

$$\Lambda_h[u]_h|_{\substack{x=mh,\\ t=n\tau}} = (a^0+a_0+a_1+a_{-1})\, u(mh, n\tau) + a^0 r h u_t(mh, n\tau) + (a_1+a_{-1})\, h u_x(mh, n\tau) +$$
$$+\frac{1}{2} a^0 r^2 h^2 u_{tt}(mh, n\tau) + \frac{1}{2}(a_1+a_{-1})\, h^2 u_{xx}(mh, n\tau) + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3). \tag{14}$$

Transformons cette expression. Déduisons l'identité

$$\frac{\partial^2 u}{\partial t^2} \equiv \frac{\partial^2 u}{\partial x^2} + (\Lambda u)_t + (\Lambda u)_x$$

résultant de la définition de Λu :

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} + \Lambda u.$$

La démonstration est donnée par la suite d'identités évidentes

$$\frac{\partial^2 u}{\partial t^2} = \left(\frac{\partial u}{\partial x} + \Lambda u\right)_t = \frac{\partial^2 u}{\partial x\, \partial t} + (\Lambda u)_t \equiv \frac{\partial}{\partial x} u_t + (\Lambda u)_t \equiv$$
$$\equiv \frac{\partial}{\partial x}\left(u_x + \Lambda u\right) + (\Lambda u)_t \equiv u_{xx} + (\Lambda u)_t + (\Lambda u)_x.$$

Utilisons ces identités pour récrire (14) sous une forme équivalente :

$$\Lambda_h[u]_h|_{\substack{x=mh,\\ t=n\tau}} = a^0 r h (\Lambda u)_m^n + \frac{1}{2} a^0 r^2 h^2 [(\Lambda u)_t +$$
$$+ (\Lambda u)_x]_m^n + (a_0 + a^0 + a_1 + a_{-1})\, u(mh, n\tau) + (a^0 r +$$
$$+ a_1 - a_{-1})\, h u_x(mh, n\tau) + \left[\frac{1}{2} a^0 r^2 + \frac{1}{2}(a_1 + a_{-1})\right] h^2 u_{xx}(mh, n\tau) +$$
$$+ O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3). \tag{15}$$

Construisons un opérateur Λ_h vérifiant la condition $\Lambda_h u = P_h \Lambda u + O(h^2)$. Introduisons les termes en Λu, $(\Lambda u)_x$, $(\Lambda u)_t$ dans l'expression de $P_h \Lambda u$, étant donné que le choix de $P_h \Lambda u$ ne dépend que de nous. Tous les termes restants

$$(a_0 + a^0 + a_{-1} + a_1)\, u(mh, n\tau),$$
$$(a^0 r + a_1 - a_{-1})\, h u_x(mh, n\tau),$$
$$\frac{a^0 r^2 + a_1 + a_{-1}}{2}\, h^2 u_{xx}(mh, n\tau),$$
$$O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)$$

figureront nécessairement dans le reste de l'égalité

$$\Lambda_h[u]_h = P_h \Lambda u + \text{reste},$$

indépendamment du choix de l'opérateur P_h. La dernière assertion se démontre par le fait qu'il y a des fonctions $u(x, t)$ telles que u, u_x, u_{xx}, Λu, $(\Lambda u)_x$, $(\Lambda u)_t$ prennent en tout point (x_0, t_0) fixe des valeurs indépendantes quelconques données à l'avance u^0, u_x^0, u_{xx}^0,

$(\Lambda u)^0$, $(\Lambda u)^0_x$, $(\Lambda u)^0_t$. Tel est par exemple le polynôme

$$P(x, t) = u^0+u^0_x(x-x_0)+[(\Lambda u)^0+u^0_x](t-t_0)+$$
$$+\frac{1}{2}u^0_{xx}(x-x_0)^2+\frac{1}{2}[u^0_{xx}+(\Lambda u)^0_x+(\Lambda u)^0_t](t-t_0)^2+$$
$$+[u^0_{xx}+(\Lambda u)^0_x](x-x_0)(t-t_0).$$

Les valeurs de u, u_x, u_{xx}, Λu, $(\Lambda u)_t$, $(\Lambda u)_x$ étant indépendantes et l'opérateur P_h quelconque, pour avoir l'approximation d'ordre 2 il faut que chaque terme du reste soit de l'ordre de h^2. Cette exigence s'écrit

$$\left.\begin{aligned} a_0+a^0+a_1+a_{-1} &= 0,\\ (a_0r+a_1+a_{-1})h &= 0,\\ (a_0r^2+a_1+a_{-1})h^2 &= 0.\end{aligned}\right\} \tag{16}$$

La solution du système (16) est définie à un facteur près. Adjoignons à ce système l'équation

$$a_0rh = 1, \tag{17}$$

exprimant une contrainte naturelle, encore que non nécessaire, sur le choix de l'opérateur P_h : le coefficient de (Λu) dans l'expression de $P_h\Lambda u$ est égal à l'unité.

Les seconds membres des égalités (6), (7) pourraient être complétés par des termes arbitraires $O_1(h^2)$, $O_2(h^2)$, $O_3(h^2)$, $O_4(h^2)$ si ce n'était notre convention qu'ils sont nuls.

En résolvant le système d'équations (16), (17) nous obtenons les coefficients a^0, a_{-1}, a_0, a_1 déjà cités :

$$a^0 = \frac{1}{rh},\quad a_0 = -\frac{1}{rh}+\frac{r}{h},\quad a_{-1} = \frac{1-r}{2h},\quad a_1 = -\frac{1+r}{2h}.$$

Pour ces valeurs des coefficients, le reste de (15)

$$\Lambda_h[u]_h = \Lambda u+\frac{rh}{2}(\Lambda u)_t+\frac{rh}{2}(\Lambda u)_x+$$
$$+O(a^0r^3h^3, a_1h^3, a_{-1}h^3) \equiv P_h\Lambda u+O(a^0r^3h^3, a_1h^3, a_{-1}h^3)$$

est évalué par

$$|O(a^0r^3h^3, a_1h^3, a_{-1}h^3)| \leqslant A(r^2h^2+h^2),$$

A étant une constante qui ne dépend que du maximum des valeurs absolues des dérivées troisièmes de la fonction $u(x, t)$. Ce fait se note également comme

$$|\Lambda_h[u]_h-P_h\varphi|^n_m \leqslant A(r^2+1)h^2.$$

Nous avons donc établi qu'à des modifications non essentielles près, parmi tous les schémas aux différences de la forme

$$L_hu^{(h)} = \begin{cases} a^0u^{n+1}_m+a_{-1}u^n_{m-1}+a_0u^n_m+a_1u^n_{m+1} = P_h\varphi|^n_m,\\ u^0_m = \psi(mh),\end{cases}$$

seul

$$L_hu^{(h)} = \begin{cases} \dfrac{u^{n+1}_m-u^n_m}{\tau}-\dfrac{u^n_{m+1}-u^n_{m-1}}{2h}-\dfrac{\tau}{2}\,\dfrac{u^n_{m+1}-2u^n_m+u^n_{m-1}}{h^2} =\\ \qquad = \left[\varphi+\dfrac{rh}{2}(\varphi_t+\varphi_x)\right]_{\substack{x=mh\\ t=n\tau}}\\ u^0_m = \psi(mh)\end{cases} \tag{17'}$$

approche le problème aux limites différentiel (4) pour la solution $u(x, t)$ à l'ordre deux en h.

Dans tous les exemples de schémas $L_h u^{(h)} = f^{(h)}$ de ce chapitre l'opérateur L_h de l'espace U_h dans l'espace F_h a été défini par des formules explicites. Or, on a souvent intérêt à examiner des schémas où L_h est décrit d'une façon plus compliquée. Notre exposé ultérieur nous réserve des problèmes où de tels schémas apparaissent bien naturellement.

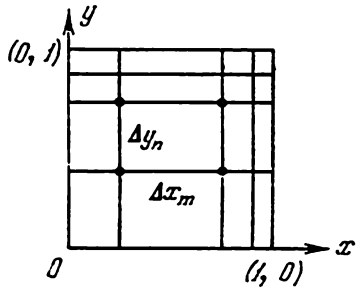


Fig. 14

Les techniques décrites de construction de schémas aux différences s'appliquent également dans le cas de problèmes à coefficients variables, de problèmes non linéaires, de réseaux à pas variable. Si l'on a par exemple le réseau irrégulier de la figure 14, on construit un schéma aux différences en remplaçant les dérivées figurant dans l'équation différentielle $u_{xx}+y_{yy} = \varphi(x, y)$ par des rapports aux différences à l'aide des formules

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{(x_m, y_n)} = \frac{\dfrac{u(x_{m+1}, y_n)-u(x_m, y_n)}{\Delta x_m} - \dfrac{u(x_m, y_n)-u(x_{m-1}, y_n)}{\Delta x_{m-1}}}{\dfrac{\Delta x_m+\Delta x_{m-1}}{2}} + \\ + \frac{1}{3}(\Delta x_m - \Delta x_{m-1})u_{xxx} + O[(\Delta x_m+\Delta x_{m-1})^2],$$

$$\left.\frac{\partial^2 u}{\partial y^2}\right|_{(x_m, y_n)} = \frac{\dfrac{u(x_m, y_{n+1})-u(x_m, y_n)}{\Delta y_n} - \dfrac{u(x_m, y_n)-u(x_m, y_{n-1})}{\Delta y_{n-1}}}{\dfrac{\Delta y_n+\Delta y_{n-1}}{2}} + \\ + \frac{1}{3}(\Delta y_n - \Delta y_{n-1}) + O((\Delta y_n+\Delta y_{n-1})^2)$$

en rejetant leurs restes.

On teste ces formules par les développements tayloriens (2). La méthode des coefficients indéterminés permet d'en établir l'unicité : il existe, à un arbitraire non essentiel près, un seul jeu de coefficients a_{-1}, a_0, a_1 pour lequel toute fonction $u(x, t)$ suffisamment régulière satisfait à la formule

$$\frac{\partial^2 u(x_m, y_n)}{\partial x^2} = a_{-1}u(x_{m-1}, y_n)+a_0u(x_m, y_n)+ \\ + a_1u(x_{m+1}, y_n)+O[\max(\Delta x_{m-1}, \Delta x_m)]$$

avec un reste infiniment petit d'ordre un en max $[\Delta x_{m-1}, \Delta x_m]$.

Une formule de la forme

$$\frac{\partial^2 u(x_m, y_n)}{\partial x^2} = a_{-1}u(x_{m-1}, y_n)+a_0u(x_m, y_n)+ \\ + a_1u(x_{m+1}, y_n)+O([\max(\Delta x_{m-1}, \Delta x_m)^2])$$

avec un reste d'ordre infinitésimal deux est inexistante pour $\Delta x_{m-1} \neq \Delta x_m$.

Pour pouvoir remplacer la dérivée par un rapport aux différences de façon plus précise, il faut posséder plus de trois points du réseau.

3. Schémas de prédiction-correction. Pour construire des schémas aux différences approchant les problèmes instationnaires on peut utiliser la même idée que celle des schémas de Runge-Kutta pour les équations différentielles ordinaires, à savoir l'idée de prédiction-correction. Cette technique permet d'élever l'ordre d'approximation obtenu avec le schéma initial. En ce qui concerne les équations différentielles quasi linéaires, la méthode en question fournit une possibilité complémentaire d'obtenir des schémas conservatifs dont nous parlerons au §30.

Rappelons l'idée de la prédiction-correction sur l'exemple d'un schéma de Runge-Kutta très simple servant à résoudre numériquement le problème de Cauchy pour l'équation différentielle ordinaire

$$\frac{dy}{dt} = f(t, y), \quad y(0) = \psi, \quad 0 < t < T. \tag{18}$$

Si l'on connaît déjà la valeur y_p au point $t_p = p\tau$, pour calculer y_{p+1} on trouve une quantité auxiliaire $\tilde{y}_{p+1/2}$ à l'aide du schéma d'Euler le plus simple (schéma prédicteur)

$$\frac{\tilde{y}_{p+1/2} - y_p}{\tau/2} = f(t_p, y_p), \tag{19}$$

et on effectue ensuite la correction par le schéma

$$\frac{y_{p+1} - y_p}{\tau} = f(t_{p+1/2}, \tilde{y}_{p+1/2}). \tag{20}$$

Disposant de la valeur prédite $\tilde{y}_{p+1/2}$ obtenue par un schéma exact à l'ordre un, on trouve approximativement le coefficient angulaire de la courbe intégrale au milieu du segment $[t_p, t_{p+1}]$ et on calcule y_{p+1} par la formule (20) avec une précision plus grande que ne l'aurait permis le schéma d'Euler (19).

Comme nous l'avons dit au n° 4 du §19, toutes ces considérations restent entières avec y, y_p, $\tilde{y}_{p+1/2}$ des vecteurs de dimension finie et f une fonction vectorielle. Il y a plus. On peut prendre y, y_p, $\tilde{y}_{p+1/2}$ pour éléments d'un espace fonctionnel et f pour opérateur dans cet espace.

Ainsi, le problème de Cauchy

$$\left.\begin{aligned} \frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} &= 0, \qquad -\infty < x < \infty, \ 0 < t < T, \\ u(x, 0) &= \psi(x), \qquad -\infty < x < \infty, \end{aligned}\right\} \tag{21}$$

$A = \text{const}$, peut être assimilé à un problème de la forme (18) en posant $y(t) = u(x, t)$ de sorte que, pour tout t, y jouera le rôle de fonction de l'argument x et l'opération f celui d'opérateur $-A\frac{\partial}{\partial x}$. Donnons un exemple de schéma aux différences prédicteur-correcteur pour le problème (21).

EXEMPLE. Soit la fonction discrète $u^p = \{u_m^p\}$, $m = 0, \pm 1, \ldots$, calculée pour p donné. Trouvons $\tilde{u}^{p+1/2} = \{\tilde{u}_{m+1/2}^{p+1/2}\}$, $m = 0, \pm 1, \ldots$, fonction auxiliaire rapportée à l'instant $t_{p+1/2} = (p+1/2)\tau$ et aux points $x_{m+1/2} = (m+1/2)h$, par le schéma exact à l'ordre un

$$\frac{\tilde{u}_{m+1/2}^{p+1/2} - \dfrac{u_{m+1}^p + u_m^p}{2}}{\tau/2} + A\,\frac{u_{m+1}^p - u_m^p}{h} = 0, \quad m = 0, \pm 1, \ldots \tag{22}$$

Employons maintenant le correcteur

$$\frac{u_m^{p+1} - u_m^p}{\tau} + A\,\frac{\tilde{u}_{m+1/2}^{p+1/2} - \tilde{u}_{m-1/2}^{p+1/2}}{h} = 0, \quad m = 0, \pm 1, \ldots \tag{23}$$

Eliminant $\tilde{u}^{p+1/2}$ entre les équations (22), (23) on obtient le schéma

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} + A\,\frac{u_{m+1}^p - u_{m-1}^p}{2h} - A^2\,\frac{\tau}{2}\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\ &u_m^0 = \psi(x_m), \quad m = 0, \pm 1, \ldots; \quad p = 0, 1, \ldots, [T/\tau] - 1 \end{aligned}\right\} \tag{24}$$

qui coïncide, lorsque $A = -1$, avec (17′). Le cas $A \neq 1$ diffère fort peu de celui considéré. Le schéma (24) et donc le schéma prédicteur-correcteur (22), (23) donnent une approximation d'ordre deux en h ; $\tau = rh$, $r = \text{const}$.

Exercices

1. Résoudre le problème de Cauchy

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = \varphi(x, t), \qquad -\infty < x < \infty, \ 0 < t < T, \\ &u(x, 0) = \psi(x), \qquad -\infty < x < \infty, \end{aligned}\right\}$$

en utilisant le réseau $x_m = mh$, $t_n = n\tau$, $h = \tau$ et en construisant un schéma aux différences de la forme

$$L_h u^{(h)} = \begin{cases} a^0 u_m^{n+1} + a^1 u_{m+1}^n + a_0 u_m^n + a_1 u_{m+1}^n = \varphi_m^n, \\ u_m^0 = \psi(mh). \end{cases}$$

Comment définir a^0, a^1, a_0, a_1 et φ_m^n pour qu'il y ait approximation d'ordre 2 en h?

2. Pour le problème de Cauchy

$$\begin{aligned} &\frac{\partial u}{\partial t} - \left(\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y}\right) = \varphi(x, y, t), \qquad -\infty < x, y < \infty, \ 0 < t < T, \\ &u(x, y, 0) = \psi(x, y), \qquad -\infty < x, y < \infty, \end{aligned}$$

se servir du réseau $x_m = mh$, $y_n = nh$, $t_p = p\tau$ et construire un schéma aux différences approximant.

3. Etant donné le problème de la chaleur

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad -\infty < x < \infty, \ 0 \leqslant t \leqslant T, \\ &u(x, 0) = \psi(x), \qquad -\infty < x < \infty, \end{aligned}\right\} \tag{25}$$

considérer pour lui le schéma aux différences

$$\left.\begin{aligned} \frac{u_m^{n+1}-u_m^n}{\tau} &= \sigma\,\frac{u_{m-1}^{n+1}-2u_m^{n+1}+u_{m+1}^{n+1}}{h^2}+(1-\sigma)\,\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2}, \\ u_m^0 &= \varphi(mh), \end{aligned}\right\}$$

avec σ un paramètre, u_m^n la valeur de la fonction cherchée au point ($x_m = mh$, $t_n = n\tau$) du réseau.

a) Montrer que, quel que soit σ, il y a approximation pour la solution régulière $u(x, t)$ à l'ordre $O(\tau+h^2)$.

b) Choisir un σ tel que l'approximation soit à l'ordre $O(\tau^2+h^2)$.

c) Relier les pas du réseau par $\tau/h^2 \equiv r = \text{const}$, puis choisir σ de façon que l'ordre de l'approximation soit h^4.

d) Le paramètre σ étant nul, choisir $r = \tau/h^2$ de façon que l'approximation soit en h^4.

e) Peut-on choisir un σ tel que pour $r = \tau/h^2$ fixe l'approximation pour toute solution régulière soit d'ordre supérieur à 4 ?

4. Pour le problème de la chaleur

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[a(x, t)\,\frac{\partial u}{\partial x}\right], \qquad -\infty < x < \infty,\ 0 < t < T,$$

$$u(x, 0) = \varphi(x), \qquad -\infty < x < \infty,$$

utiliser le réseau $x_m = mh$, $t_n = n\tau$ et construire le schéma aux différences approximant.

5. Pour le problème non linéaire de la chaleur

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[a(u)\,\frac{\partial u}{\partial x}\right], \qquad -\infty < x < \infty,\ 0 < t < T,$$

$$u(x, 0) = \varphi(x), \qquad -\infty < x < \infty,$$

utiliser le réseau $x_m = mh$, $t_n = n\tau$ et construire le schéma aux différences approximant sous forme explicite. Ecrire les formules de calcul de $u^{(k)}$ par ce schéma.

6. Démontrer qu'étant donnée une fonction discrète bornée $u^p = \{u_m^p\}$, il y a existence et unicité d'une fonction discrète bornée $u^{p+1} = \{u_m^{p+1}\}$ définie par le schéma aux différences

$$\frac{u_m^{p+1}-u_m^p}{\tau} - \frac{u_{m+1}^{p+1}-u_{m-1}^{p+1}}{2h} = 0, \qquad m = 0, \pm 1, \ldots$$

7. Démontrer que le schéma prédicteur-correcteur pour le problème (25) où les valeurs de la solution $\{\tilde{u}_m^{p+1/2}\}$ à un niveau intermédiaire se définissent par le schéma implicite approchant à l'ordre $O(\tau+h^2)$

$$\frac{\tilde{u}_m^{p+1/2}-u_m^p}{\tau/2} - \frac{\tilde{u}_{m+1}^{p+1/2}-2\tilde{u}_m^{p+1/2}+\tilde{u}_{m-1}^{p+1/2}}{h^2} = 0, \qquad m = 0, \pm 1, \ldots,$$

et la solution $\{u_m^{p+1}\}$ se définit par

$$\frac{u_m^{p+1}-u_m^p}{\tau} - \frac{\tilde{u}_{m+1}^{p+1/2}-2\tilde{u}_m^{p+1/2}+u_{m-1}^{p+1/2}}{h^2} = 0, \qquad u_m^0 = \varphi(x_m), \qquad m = 0, \pm 1, \ldots,$$

donne une approximation en $O(\tau^2+h^2)$ pour la solution régulière u.

§23. Exemples de construction de conditions aux limites

En sélectionnant les exemples à discuter au §22 nous l'avons fait de manière qu'aucune question ne se pose quant à la construction de conditions aux limites aux différences : on les a obtenues sans peine à partir

de celles différentielles, et ce de façon que les équations aux différences soient vérifiées exactement par la substitution de $[u]_h$. Dans le présent paragraphe nous allons examiner des exemples plus compliqués dans ce sens.

EXEMPLE 1. En construisant un schéma aux différences pour le problème

$$Lu = \begin{cases} u_t - u_x = \varphi(x, t), \\ u(x, 0) = \psi(x) \end{cases} \tag{1}$$

utilisons l'équation aux différences

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} - \frac{u_{m+1}^n - u_{m-1}^n}{2h} = \varphi(mh, n\tau), \tag{2}$$

$$n = 1, 2, \ldots; \quad m = 0, \pm 1, \ldots; \quad \tau = rh.$$

La solution de celle-ci se calcule en se donnant, en plus de u_m^0,

$$u_m^0 = \psi(mh), \quad m = 0, \pm 1, \ldots, \tag{3}$$

u_m^1, $m = 0, \pm 1, \ldots$ Moyennant (2) on trouve alors, pour $n = 1, 2, \ldots$, u_m^2, $m = 0, \pm 1, \ldots$, puis u_m^3, $m = 0, \pm 1, \ldots$, et ainsi de suite. Les valeurs données u_m^1 doivent être voisines de

$$u(mh, \tau) = u(mh, 0) + \tau u_t(mh, 0) + O(\tau^2).$$

Puisque $u_t = u_x + Au$, $Au \equiv u_t - u_x = \varphi(x, t)$, $u(x, 0) = \psi(x)$, on a

$$u(mh, \tau) = u(mh, 0) + \tau[u_x + Au]_{\substack{x=mh, \\ t=0}} + O(\tau^2) =$$

$$= \psi(mh) + \tau[\psi'(mh) + \varphi(mh, 0)] + O(\tau^2).$$

En rejetant le terme $O(\tau^2)$ on pose donc

$$u_m^1 = \psi(mh) + \tau[\psi'(mh) + \varphi(mh, 0)]. \tag{4}$$

Il est clair que l'approximation du problème aux limites différentiel (1) par le schéma aux différences

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{n+1} - u_m^{n-1}}{2\tau} - \dfrac{u_{m+1}^n - u_{m-1}^n}{2h} = \varphi(mh, n\tau), \\ u_m^0 = \psi(mh), \\ u_m^1 = \psi(mh) + \tau[\psi'(mh) + \varphi(mh, 0)] \end{cases} \tag{5}$$

est d'ordre 2 en h. Ce schéma est non trivial: l'équation auex différences (2) est d'ordre deux en t et l'équation différentielle (1) d'ordre un. Aussi on a eu à construire la seconde condition aux limites aux différences (4) qui ne découle pas directement de la condition aux limites donnée pour le problème différentiel.

Voici un autre exemple où la construction de conditions aux limites aux différences n'est pas évidente.

EXEMPLE 2. Soit le problème aux limites différentiel

$$Lu = \begin{cases} u_t - u_x = \varphi(x, t), & 0 < x < 1, 0 < t < T, \\ u(0, x) = \psi_0(x), & 0 < x < 1, \\ u(t, 1) = \psi_1(t), & 0 < t < T. \end{cases} \tag{6}$$

Toute solution de l'équation différentielle du problème (6) est définie de façon unique connaissant sa valeur en un point de chaque droite $x+t = \text{const}$. En effet, le long d'une telle droite

$$\frac{du}{dt} = u_t + u_x \frac{dx}{dt} = u_t - u_x = \varphi(x, t),$$

si bien que $u(x, t)$ est une intégrale de $\varphi(x, t)$ relative à la droite

$$x+t = \text{const}.$$

La valeur de la constante d'intégration est fonction de celle de u au point donné.

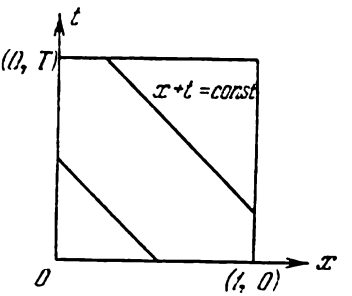


Fig. 15

La figure 15 représente le rectangle $0 \leqslant x \leqslant 1$, $0 \leqslant t \leqslant T$ dans lequel nous nous proposons de chercher la solution, et la famille de droites parallèles $x+t = \text{const}$ dont chacune coupe en un point soit le segment $0 \leqslant x \leqslant 1$ de l'axe Ox, soit le segment $0 \leqslant t \leqslant T$ de la droite $x = 1$, qui est l'intervalle de définition de $u(x, t)$. Ainsi, le problème (6) admet une solution unique.

Construisons le schéma aux différences pour la résolution numérique de (6). Donnons-nous un h tel que $Mh = 1$ et posons $\tau = rh$, M étant un entier et $r = \text{const}$. Formons un réseau D_h avec les points $(mh, n\tau)$, $m = 0, 1, \ldots M$; $n = 0, 1, \ldots, [T/\tau]$.

Faisons correspondre aux points de D_h autres que ceux de la frontière supérieure et des bords du rectangle D l'équation

$$\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_{m-1}^n}{2h} - \frac{\tau}{2} \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = \varphi_m^n, \tag{7}$$

où

$$\varphi_m^n = \left[\varphi(x, t) + \frac{rh}{2}(\varphi_t + \varphi_x)\right]_{\substack{x=mh \\ t=n\tau}}. \tag{8}$$

La façon dont nous l'avons obtenue est décrite en détail au §22.

Les valeurs u_m^0 et u_M^n sont données par les égalités

$$\left.\begin{aligned} u_m^0 &= \psi_0(mh), & m &= 0, 1, \ldots, M-1, \\ u_M^n &= \psi_1(n\tau), & n &= 0, 1, \ldots, N, \; N \equiv [T/\tau], \end{aligned}\right\} \tag{9}$$

analogues aux conditions aux limites pour le problème différentiel considéré. Les égalités (9) ne suffisent cependant pas pour définir la solution u_m^n partout sur D_h, les valeurs u_0^{n+1} sur le bord de gauche restant indé-

terminées. On complète donc les conditions aux limites aux différences par

$$\frac{u_0^{n+1}-u_0^n}{\tau}-\frac{u_1^n-u_0^n}{h}=\varphi(0, n\tau), n=0, 1, \ldots, N-1, \tag{10}$$

condition qui apparaît en remplaçant dans l'égalité

$$\frac{\partial u(0, t)}{\partial t}-\frac{\partial u(x, 0)}{\partial x}=\varphi(0, t),$$

résultant de l'équation différentielle donnée (6), les dérivées par des rapports aux différences correspondants.

Ainsi, nous avons formé le schéma aux différences $L_h u^{(h)}=f^{(h)}$:

$$L_h u^{(h)}=\begin{cases} \dfrac{u_m^{n+1}-u_m^n}{\tau}-\dfrac{u_{m+1}^n-u_{m-1}^n}{2h}-\dfrac{\tau}{2}\,\dfrac{u_{m+1}^n-2u_m^n+u_{m-1}^n}{h^2}, \\ \qquad m=1, 2, \ldots, M-1\,; \quad n=0, 1, \ldots, N-1, \\ u_m^0,\ m=0, 1, \ldots, M-1, \\ u_M^n,\ n=0, 1, \ldots, N, \\ \dfrac{u_0^{n+1}-u_0^n}{\tau}-\dfrac{u_1^n-u_0^n}{h}, \quad n=0, 1, \ldots, N-1, \end{cases}$$

$$f^{(h)}=\begin{cases} \left[\varphi+\dfrac{rh}{2}(\varphi_t+\varphi_x)\right]_{\substack{x=mh,\\ t=n\tau}}, \\ \psi_0(mh),\ m=0, 1, \ldots, M-1, \\ \psi_1(n\tau),\ \ n=0, 1, \ldots, N, \\ \varphi(0, n\tau),\ n=0, 1, \ldots, N-1. \end{cases}$$

Etablissons l'ordre d'approximation. A la lumière des raisonnements du §22, on constate que sous l'hypothèse de régularité suffisante de la solution $u(x, t)$ le résidu $\delta f^{(h)}$ dû à la substitution de $[u]_h$ dans le schéma aux différences, $L_h[u]_h=f^{(h)}+\delta f^{(h)}$, est de la forme :

$$\delta f^{(h)}\equiv\begin{cases} O_{mn}(h^2)=O(h^2), & m=1, 2, \ldots, M-1; \\ & n=0, 1, \ldots, N-1, \\ 0, & m=0, 1, \ldots, M-1, \\ 0, & n=0, 1, \ldots, N, \\ \dfrac{\tau}{2}u_{tt}(0, n\tau+\xi_1\tau)+\dfrac{h}{2}u_{xx}(\xi_2 h, n\tau), & n=0, 1, \ldots, N-1, \\ & 0<\xi_1<1, \quad 0<\xi_2<1. \end{cases}$$

Si F_h est muni d'une norme en posant

$$g^{(h)}=\begin{cases} A_m^n, \\ a_m, \\ b^n, \\ c^n, \end{cases} \qquad \|g^{(h)}\|_{F_h}=\max_{m,n}|A_m^n|+\max_m|a_m|+\max_n|b^n|+\max_n|c^n|$$

pour un $g^{(h)} \in F_h$ arbitraire, alors $\|\delta f^{(h)}\|_{F_h} = O(h)$ et l'approximation ne sera que d'ordre un en h. L'expression de $\delta f^{(h)}$ montre que la cause en est le résidu $\frac{\tau}{2}u_{tt}+\frac{h}{2}u_{xx} = O(h)$, apparaissant par substitution de $[u]_h$ dans la condition aux limites complémentaire au bord de gauche que nous avons construite artificiellement.

Le reste pour la norme $\|\cdot\|_{F_h}$ utilisée est évalué par les seules dérivées secondes de la solution, i.e. cette norme ne permet pas de se servir, lorsqu'on étudie les conditions aux limites, de la régularité de la solution qui était nécessaire pour obtenir l'approximation d'ordre 2 aux points intérieurs.

Donnons la norme $\|\cdot\|_{F_h}$ pour laquelle le schéma aux différences construit plus haut réalise l'approximation d'ordre 2 pour $u(x, t)$ suffisamment régulière :

$$\|g^{(h)}\|_{F_h} = h \max_n |c^n| + h \max_n \left|\frac{c^{n+1}-c^n}{\tau}\right| + \left(h \sum_{m=0}^{N} |a_m|^2\right)^{1/2} + \\ + \max_n |b^n| + \max_n \left|\frac{b^{n+1}-b^n}{\tau}\right| + \max_{m,n} |A_m^n|.$$

On voit aisément que dans le cas de ce schéma

$$\|\delta f^{(h)}\|_{F_h} \leqslant A(rh^2+h^2), \quad r = \tau/h,$$

et la constante A s'évalue par des dérivées d'ordre 3 inclus de $u(x, t)$.

La régularité dans cette norme est repérée par les termes

$$\left|\frac{c^{n+1}-c^n}{\tau}\right|, \quad \left|\frac{b^{n+1}-b^n}{\tau}\right|.$$

Le lecteur a certes remarqué que certains termes de la formule donnant une nouvelle norme dans F_h diffèrent des termes correspondants dans la première norme par le facteur h. On conçoit qu'en multipliant arbitrairement par h et par diverses puissances de h on aboutit à une approximation de tout ordre voulu. Mais nous avons déjà discuté le choix des normes dans le cas des équations différentielles ordinaires (§13) et nous savons que seules sont raisonnables les normes pour lesquelles le schéma aux différences approche le problème aux limites différentiel tout en étant stable.

Au §40 nous démontrerons la stabilité du schéma considéré avec la norme utilisée relativement à laquelle il y a approximation d'ordre 2.

L'exemple 2 est extrêmement instructif. Car il montre qu'une vérification raisonnable de l'approximation nécessite un choix heureux de la norme. En étudiant un schéma on essaie pas mal de normes et il faut dans chaque cas analyser la stabilité, ce qui exige souvent, au moins aujourd'hui, de grands efforts et beaucoup d'ingéniosité.

Dans la pratique, on remplace dans la plupart des cas le problème à étudier par un problème modèle simplifié, puis on contrôle en calculant selon le schéma aux différences pour le problème originel.

Exercices

1. Soit le problème de Cauchy

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = \varphi(x, t), \qquad -\infty < x < \infty, \qquad 0 \leqslant t \leqslant T,$$

$$u(x, 0) = \psi_1(x), \qquad -\infty < x < \infty,$$

$$\frac{\partial u(x, 0)}{\partial t} = \psi_2(x), \qquad -\infty < x < \infty.$$

Etudier l'approximation réalisée pour $u(x, t)$ suffisamment régulière par le schéma aux différences

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{\tau^2} - \dfrac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = \varphi(mh, n\tau), \\ u_m^0 = \psi_1(mh), \\ \dfrac{u_m^1 - u_m^0}{\tau} = (\psi_2)_m \end{cases}$$

si $(\psi_2)_m = \psi_2(mh)$. Accepter pour norme $\|f^{(h)}\|_{F_h}$ le maximum des modules de toutes les composantes de l'élément

$$f^{(h)} = \begin{cases} \varphi_m^n, \\ \psi_1(mh), \\ (\psi_2)_m. \end{cases}$$

Montrer que l'approximation est du premier ordre en h ; $\tau = rh$, $r = \text{const}$.

Etant donnés $\varphi(x, t)$, $\psi_1(x)$ et $\psi_2(x)$, quels doivent être $(\psi_2)_m$ pour que l'approximation soit d'ordre 2 ?

2. Pour le problème de propagation de la chaleur sur un segment

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = \varphi(x, t), \qquad 0 < x < 1, \qquad 0 < t < T,$$

$$u(x, 0) = \psi_0(x), \qquad 0 < x < 1,$$

$$\frac{\partial u(0, t)}{\partial x} = \psi_1(t),$$

$$0 < t < T,$$

$$u(1, t) = \psi_2(t),$$

considérer un schéma aux différences de la forme

$$\left.\begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2} = \varphi(mh, n\tau), \\ &m = 1, 2, \ldots, M-1; \qquad n = 0, 1, \ldots, [T/\tau]-1, \\ &u_m^0 = \psi_0(mh), \qquad m = 0, 1, \ldots, M, \\ &\frac{u_1^n - u_0^n}{h} = (\psi_1)^n, \qquad n = 1, 2, \ldots, [T/\tau], \\ &u_M^n = \psi_2(n\tau), \qquad n = 1, 2, \ldots, [T/\tau]. \end{aligned}\right\}$$

Prendre pour norme $\|\cdot\|_{F_h}$ le maximum des valeurs absolues des seconds membres

des équations constituant dans l'ensemble le schéma aux différences en question. Les pas h et τ sont supposés reliés par $\tau = rh^2$, $r = \text{const}$. Montrer qu'en posant $(\psi_1)^n = \psi_1(nh)$ on aboutit à un schéma approchant à l'ordre un pour une solution régulière. Donner la formule définissant $(\psi_1)^n$ tel que l'ordre d'approximation soit 2.

§24. Condition nécessaire de convergence de Courant, Friedrichs et Lewy

Nous avons démontré au §21 que le schéma aux différences

$$\left.\begin{aligned} \frac{u_m^{n+1}-u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} &= 0, \\ u_m^0 &= \psi(mh) \end{aligned}\right\} \tag{1}$$

approchant le problème de Cauchy

$$\left.\begin{aligned} \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} &= 0, \quad 0 < t < T, \\ u(x, 0) &= \psi(x) \end{aligned}\right\} \tag{2}$$

ne peut être convergent pour $\psi(x)$ arbitraire si $\tau/h > 1$ (voir fig. 9, p. 162). Nous avons utilisé en l'occurrence une considération générale énoncée pour la première fois, pour un exemple quelque peu différent, par Courant, Friedrichs et Lewy.

Cette considération aide souvent à construire et à étudier des schémas aux différences.

1. Condition de Courant, Friedrichs et Lewy. Admettons que la position d'un problème différentiel implique une fonction ψ (voir, par exemple, (2)). Choisissons un point quelconque P du domaine de définition de la solution u et supposons que la valeur $u(P)$ de la solution dépend des valeurs de ψ aux points d'un ensemble $G_\psi = G_\psi(P)$ appartenant au demaine de définition de ψ, i.e. en faisant varier ψ dans un voisinage restreint de tout point Q de $G_\psi(P)$ on modifie la valeur $u(P)$. Soit $L_h u^{(h)} = f^{(h)}$ un schéma aux différences utilisé pour calculer la solution u, les valeurs de la solution $u^{(h)}$ au point du réseau le plus proche de P étant complètement définies par celles de ψ sur un ensemble $G_\psi^{(h)} = G_\psi^{(h)}(P)$.

Pour qu'il y ait convergence $u^{(h)} \to u$ lorsque $h \to 0$, le schéma aux différences doit être tel que pour $h \to 0$ un voisinage quelconque de tout point du domaine $G_\psi(P)$ contienne, pour h suffisamment petit, un point de l'ensemble $G_\psi^{(h)} = G_\psi^{(h)}(P)$.

Supposons la condition énoncée de Courant, Friedrichs et Lewy non remplie, de sorte qu'un certain voisinage fixe d'un point Q du domaine $G_\psi(P)$ ne renferme pas, pour tous les h suffisamment petits, de points de $G_\psi^{(h)} = G_\psi^{(h)}(P)$. Si la convergence $u^{(h)} \to u$ a (par hasard !) lieu pour la fonction ψ donnée, on modifie ψ dans le voisinage indiqué de Q de façon que $u(P)$ varie, en conservant ψ partout ailleurs. Pour cette nouvelle fonction ψ on n'a plus la convergence $u^{(h)} \to u$: la valeur $u(P)$ a changé et les valeurs

de $u^{(h)}$ au point le plus proche de P se sont conservées pour h petits puisque la fonction ψ reste la même aux points de l'ensemble $G_\psi^{(h)} = G_\psi^{(h)}(P)$.

La condition de Courant, Friedrichs et Lewy se met facilement sous forme d'un théorème se démontrant moyennant les considérations ci-dessus. Nous préférons cependant de nous en abstenir.

Examinons plusieurs exemples où la considération exposée permet d'établir la divergence et l'impraticabilité d'un schéma aux différences et de mettre en évidence un autre qui soit stable et convergent. La démonstration de la convergence est sans doute un processus autonome, car la condition de Courant, Friedrichs et Lewy est nécessaire mais non suffisante pour avoir la convergence. Notons de plus qu'en présence d'approximation il s'agit également d'une condition nécessaire de stabilité vu que la convergence découle de l'approximation et de la stabilité.

2. Exemples de schémas aux différences pour le problème de Cauchy. Utilisons la condition de Courant, Friedrichs et Lewy pour analyser certains schémas aux différences approchant le problème de Cauchy

$$\left.\begin{aligned} \frac{\partial u}{\partial t}+a(t)\frac{\partial u}{\partial x} &= \psi_0(x, t), \qquad -\infty < x < \infty,\ 0 < t < \tau, \\ u(x, 0) &= \psi_1(x), \qquad -\infty < x < \infty, \end{aligned}\right\} \tag{3}$$

où $\psi_0(x, t)$ et $\psi_1(x)$ sont les « entrées » données du problème (3) et

$$a(t) \equiv -1-2t.$$

La solution de (3) en un point (x_P, t_P) dépend des valeurs des fonctions $\psi_0(x, t)$ et $\psi_1(x)$ en tous les points que traverse la caractéristique de l'équation différentielle (3) qui part d'un point A de l'axe Ox et aboutit au point P.

En effet, les caractéristiques sont ici les courbes intégrales de l'équation différentielle

$$\frac{dx}{dt} = a(t),$$

i.e. les paraboles $x = -t^2-t+C$. On a le long de chaque caractéristique

$$\frac{du}{dt} = \frac{\partial u}{\partial t}+\frac{\partial u}{\partial x}\frac{dx}{dt} = \frac{\partial u}{\partial t}+a(t)\frac{\partial u}{\partial x} = \psi_0(x, t).$$

La valeur $u(x_P, t_P)$ de la solution en un point $P = (x_P, t_P)$ s'exprime donc par la formule

$$u(x_P, t_P) = \psi_1(A)+\int_0^{t_P} \psi_0[x(t), t]\,dt = \psi_1(A)+\int_{AQP} \psi_0(x, t)\,dt$$

avec A un point de l'axe Ox et AQP un morceau de la caractéristique.

La figure 16 représente la caractéristique $x = 2-t-t^2$ issue du point $A = (2, 0)$ et aboutissant au point $P = (0, 1)$. Nous voyons que la valeur $u(P) \equiv u(x_P, t_P)$ de la solution du problème (3) dépend de la valeur de

$\psi_1(x)$ au point A si bien que $A = G_{\psi_1}(P)$. Ensuite, $u(P)$ dépend des valeurs de $\psi_0(x, t)$ sur le morceau AQP de la caractéristique. AQP est bien $G_{\psi_0}(P)$.

Considérons le schéma aux différences

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} + a(t_n)\dfrac{u_m^n - u_{m-1}^n}{h} = \psi_0(x_m, t_n), \\ \qquad\qquad\qquad\qquad\qquad\qquad u_m^0 = \psi_1(x_m), \\ m = 0, \pm 1, \ldots, \qquad\qquad n = 0, 1, \ldots, [1/\tau] - 1, \end{cases} \tag{4}$$

ou

$$\left.\begin{aligned} u_m^{n+1} &= [1 + a(t_n)r]\, u_m^n - a(t_n)\, r u_{m-1}^n + \tau\psi_0(x_m, t_n), \\ u_m^0 &= \psi_1(x_m), \end{aligned}\right\} \tag{5}$$

où $x_m = mh$, $t_n = n\tau$, $r = \tau/h$, $a(t) \equiv -1 - 2t$. Montrons que ce schéma ne converge pour aucun rapport r des pas vu qu'il ne satisfait pour aucun r la condition de Courant, Friedrichs et Lewy.

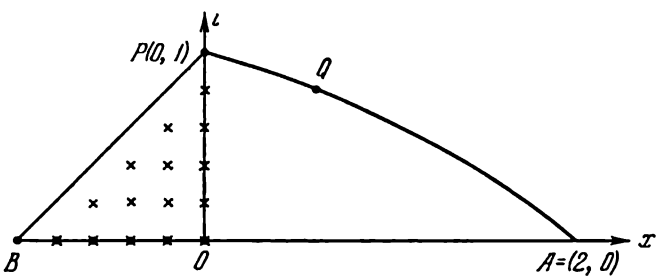


Fig. 16

Acceptons pour P le point (0, 1) et choisissons un réseau tel que $N\tau = 1$. La valeur $u^{(h)} = u^{(h)}(P)$ de la solution au point $P = (0, 1)$, i.e. u_0^N, s'exprime, en vertu de la formule aux différences (5), par la valeur $\psi_0(0, 1-\tau)$ et par les valeurs u_{-1}^{N-1}, u_0^{N-1} qui sont représentées à leur tour moyennant $\psi_0(-h, 1-2\tau)$, $\psi_0(0, 1-2\tau)$ et les trois valeurs u_{-2}^{N-2}, u_{-1}^{N-2}, u_0^{N-2}, et ainsi de suite. La valeur finale u_0^N sera donnée par les valeurs de la fonction $\psi_0(x, t)$ aux points du réseau repérés par les petites croix (fig. 16) et les valeurs $u_{-N}^0 = \psi_1(x_{-N})$, $u_{-N+1}^0 = \psi_1(x_{-N+1})$, $\ldots$, $u_0^0 = \psi_1(x_0)$ que la fonction $\psi_1(x)$ prend aux points x_{-N}, x_{-N+1}, $\ldots$, x_0 de l'axe Ox. Ainsi, l'ensemble $G_{\psi_0}^{(h)}(P)$ est celui des points marqués par les croix et $G_{\psi_1}^{(h)}(P)$ celui des points x_{-N}, x_{-N+1}, $\ldots$, x_0 de l'axe Ox (ces ensembles ont des points communs, à savoir les points de l'axe Ox). Il est évident que tout point Q de $G_{\psi_0}(P)$ possède un voisinage ne contenant pas les points de $G_{\psi_0}^{(h)}(P)$, quelque petit que soit h. Le schéma aux différences (4) ne vérifie pas la condition nécessaire de convergence de Courant, Friedrichs et Lewy.

Prenons maintenant pour le problème (3) le schéma aux différences (fig. 17)

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1}-u_m^n}{\tau} + a(t_n)\dfrac{u_{m+1}^n - u_m^n}{h} = \psi_0(x_m, t_n), \\ \qquad\qquad\qquad\qquad u_m^0 = \psi_1(x_m), \\ m = 0, \pm 1, \ldots; \qquad n = 0, 1, \ldots, 1/\tau - 1, \end{cases} \tag{6}$$

ou

$$\left.\begin{aligned} u_m^{n+1} &= [1+a(t_n)r]\,u_m^n - a(t_n)\,r u_{m+1}^n + \tau\psi_0(x_m, t_n), \\ u_m^0 &= \psi_1(x_m), \end{aligned}\right\} \tag{7}$$

avec $r = \tau/h$.

Le pas τ du réseau est choisi à partir de la condition $N\tau = 1$, N entier, de sorte que le point $P = (0, 1)$ appartienne au réseau. En vertu des formules (7), la valeur de la solution $u^{(h)}$ en ce point, i.e. u_0^N, s'exprime par l'intermédiaire de $\psi_0(0, 1-\tau)$ et de u_0^{N-1} et u_1^{N-1} et celles-ci par $\psi_0(0, 1-2\tau)$, $\psi_0(h, 1-2\tau)$ et trois valeurs $u_0^{N-2}, u_1^{N-2}, u_2^{N-2}$ et ainsi de suite. Finalement u_0^N est donné par les valeurs de $\psi_0(x_m, t_m)$ aux points du réseau marqués par les petites croix sur la figure 17 et les valeurs $u_0^0 = \psi_1(0)$, $u_1^0 = \psi_1(x_1)$, ...

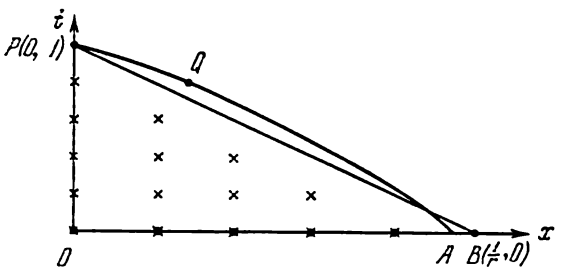


Fig. 17

..., $u_N^0 = \psi_1(x_N)$ de $\psi_1(x)$ aux points $x_0, x_1, \ldots, x_N$ de l'axe Ox. Ainsi, $G_{\psi_0}^{(h)}(P)$ est dans ce cas l'ensemble de points repérés par les croix et $G_{\psi_1}^{(h)}(P)$ celui des points $x_0, x_1, \ldots, x_N$ de l'axe Ox. Il est clair que si $r = \tau/h > 1/2$ (la figure ne représente pas ce cas), le point $B = (1/r, 0)$ est à gauche du point $A = G_{\psi_1}(P)$ et il existe donc un voisinage de A qui ne contient pas pour $h \to 0$ les points de $G_{\psi_1}^{(h)}(P)$. La condition de Courant, Friedrichs et Lewy étant violée, il n'y a pas lieu de s'attendre à la convergence.

Pour que le schéma (6) puisse converger, il faut que $r \leqslant 1/2$. Cela ne suffit cependant pas. Supposons que $r < 1$ et qu'un point Q de la caractéristique AQP se place au-dessus de la droite BP, comme on le voit fig. 17. Dans ce cas, on n'aura pas non plus de convergence. La valeur de $\psi_0(x, t)$ au point Q influe sur la valeur $u(0, 1)$ de la solution du problème différentiel, i.e. $Q \in G_{\psi_0}(P)$. Mais cette valeur de $\psi_0(x, t)$ (pas plus que ses valeurs sur

tout le morceau QP de la caractéristique) n'exerce aucune influence sur la valeur $u^{(h)}(P)$ de la solution de l'équation aux différences au point P: il existe un voisinage de Q qui ne contient pas pour $h \to 0$ de points de l'ensemble $G_{\varphi_0}^{(h)}(P)$. Conclusion : la condition de Courant, Friedrichs et Lewy n'est pas remplie.

A condition de prendre r si petit que le triangle OPB contient non seulement le point $A = (2, 0)$, mais aussi toute la caractéristique AQP, on démontre la stabilité (et la convergence) du schéma aux différences (6). On effectue ce choix de r compte tenu du fait que (en vertu de l'équation différentielle de la caractéristique $dx/dt = a(t)$) la quantité $-1/a(t)$ est la pente de la tangente à la caractéristique et $-r = -\tau/h$ celle de la droite BP. On conçoit donc que la caractéristique AQP est dans le triangle BOP si

$$r \leqslant \frac{1}{\max\limits_{0 \leqslant t \leqslant 1} |a(t)|} = \frac{1}{3}, \quad \tau \leqslant \frac{h}{3}, \tag{8}$$

et on a la condition de Courant, Friedrichs et Lewy.

Montrons que dans la condition (8) le schéma aux différences (6) approchant le problème de Cauchy (3) est stable et donc convergent. Les normes seront définies par les égalités

$$\|u^{(h)}\|_{U_h} = \max_{m,n} |u_m^n|,$$

$$\|f^{(h)}\|_{F_h} = \max_{m,n} |\psi_0(x_m, t_n)| + \max_m |\psi_1(x_m)|.$$

Vu que dans la condition (8)

$$1 + a(t_n)r \geqslant 1 - \frac{2t_n+1}{3} \geqslant 0, \quad 0 \leqslant t_n \leqslant 1,$$

l'égalité (7) fournit

$$|u_m^{n+1}| \leqslant \left[1 - \frac{2t_n+1}{3} r + \frac{2t_n+1}{3} r\right] \max_m |u_m^n| + \tau \max_{m,n} |\psi_0(x_m, t_n)| \leqslant$$

$$\leqslant \max_m |u_m^n| + \tau \max_{m,n} |\psi_0(x_m, t_n)| \leqslant \max_m |u_m^{n-1}| + 2\tau \max_{m,n} |\psi_0(x_m, t_n)| \leqslant$$

. . . . . .

$$\leqslant \max_m |u_m^0| + (n+1)\tau \max_{m,n} |\psi_0(x_m, t_n)| \leqslant$$

$$\leqslant \max_m |\psi_1(x_m)| + 1 \cdot \max_{m,n} |\psi_0(x_m, t_n)| = \|f^{(h)}\|_{F_h}.$$

Puisque l'inégalité obtenue

$$|u_m^{n+1}| \leqslant \|f^{(h)}\|_{F_h}$$

est juste quels que soient $m = 0, \pm 1, \ldots,$ et n, $(n+1)\tau \leqslant 1$, on a

$$\|u^{(h)}\|_{U_h} \leqslant \|f^{(h)}\|_{F_h},$$

ce qui démontre la stabilité du schéma (6) sous l'hypothèse (8). On peut

affaiblir la contrainte (8) imposée au pas τ pour h donné, $\tau \leqslant {}^1/_3 h$, sans violer la condition de Courant, Friedrichs et Lewy si l'on rend τ variable, $t_{n+1} = t_n + \tau_n$, et si on le choisit pour le passage $t_n \to t_{n+1}$ compte tenu de la pente de la caractéristique au voisinage du point $t = t_n$, notamment à partir de la condition

$$r_n \equiv \frac{\tau_n}{h} \leqslant \frac{1}{|a(t_n)|} = \frac{1}{2t_n+1}, \quad n = 0, 1, \ldots \tag{9}$$

Le schéma (6) ainsi modifié se présente comme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau_n} + a(t_n) \dfrac{u_{m+1}^n - u_m^n}{h} = \psi_0(x_m, t_n), \\ u_m^0 = \psi_1(x_m) \end{cases} \tag{10}$$

ou

$$\left.\begin{aligned} u_m^{n+1} &= [1 + a(t_n) r_n]\, u_m^n - a(t_n) r_n u_{m+1}^n + \tau_n \psi_0(x_m, t_n), \\ u_m^0 &= \psi_1(x_m). \end{aligned}\right\} \tag{11}$$

Conformément à la formule (9) la contrainte sur le pas τ_n est moins forte que pour le schéma (6) à pas constant. Pour n petits, on utilise le pas $\tau_n \approx h$, et ce n'est que lorsque t_n avoisine $t = 1$, qu'on choisit $\tau_n = {}^1/_3 h$ (fig. 18). La démonstration de la stabilité du schéma (10) dans la condition

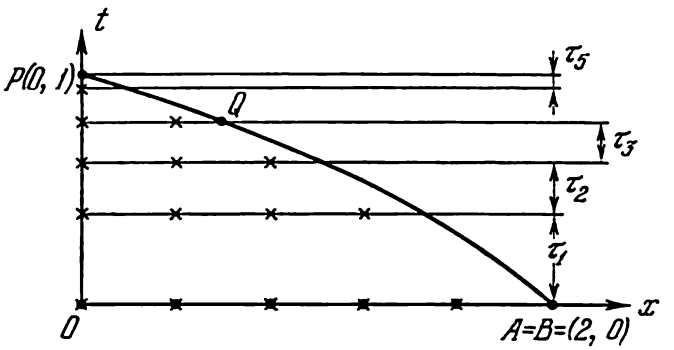


Fig. 18

(9) ne diffère que très peu de celle de la stabilité de (6) dans la condition (8) : utilisant l'inégalité $1 + a(t_n)\, r_n \geqslant 0$ on obtient en vertu de (11)

$$\begin{aligned} |u_m^{n+1}| &\leqslant \max_m |u_m^n| + \tau_n \max_{m,n} |\psi_0(x_m, t_n)| \leqslant \max_m |u_m^{n-1}| + \\ &+ (\tau_{n-1} + \tau_n) \max_{m,n} |\psi_0(x_m, t_n)| \leqslant \max_m |u_m^0| + \\ &+ t_{n+1} \max_{m,n} |\psi_0(x_m, t_n)| \leqslant \| f^{(h)} \|_{F_h}. \end{aligned}$$

D'où

$$\| u^{(h)} \|_{U_h} \leqslant \| f^{(h)} \|_{F_h}$$

exprimant la stabilité.

3. Exemples de schémas aux différences pour le problème de Dirichlet. Utilisons la condition de Courant, Friedrichs et Lewy afin d'analyser deux schémas aux différences approchant le problème de Dirichlet relatif à l'équation de Poisson :

$$\left.\begin{aligned} \frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2} &= \varphi(x, y) \quad 0 \leqslant x, y \leqslant 1, \\ u|_\Gamma &= \psi(x, y), \quad (x, y) \in \Gamma, \end{aligned}\right\} \tag{12}$$

dans le carré $D = (0 \leqslant x, y \leqslant 1)$ de frontière Γ. Construisons le réseau $x_m = mh$, $y_n = nh$, où $h = 1/M$, M entier (fig. 19,*a*). Le réseau D_h com-

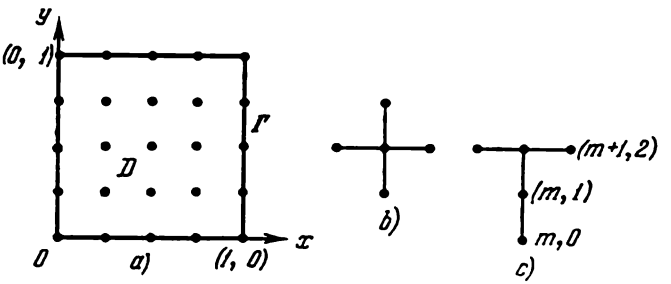


Fig. 19

prendra les points (x_m, y_n) intérieurs au carré D ou situés sur son contour. Considérons le schéma aux différences approchant le problème (12) :

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{m+1,n}-2u_{mn}+u_{m-1,n}}{h^2}+\dfrac{u_{m,n+1}-2u_{mn}+u_{m,n-1}}{h^2} = \\ \qquad = \varphi(mh, nh) \text{ si } (mh, nh) \in D \\ u_{mn} = \psi(mh, nh) \text{ si } (mh, nh) \in \Gamma. \end{cases} \tag{13}$$

Ce dernier s'obtient en remplaçant les dérivées u_{xx} et u_{yy} par des rapports aux différences, et l'approximation a bien lieu. Au §34 nous démontrerons la stabilité de (13) et aux trois paragraphes suivants nous exposerons les procédés de calcul de la solution $u^{(h)}$. Notons que ce calcul n'est pas simple étant donné une complexité suffisante du système d'équations $L_h u^{(h)} = f^{(h)}$ servant à déterminer les valeurs de $u^{(h)}$ pour des h faibles. Cette complexité seule incite à construire un schéma dont les solutions se calculent de façon élémentaire. De prime abord le schéma

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{m-1,n}-2u_{mn}+u_{m+1,n}}{h^2}+\dfrac{u_{m,n+1}-2u_{mn}+u_{m,n-1}}{h^2} = \varphi(mh, nh), \\ \qquad m = 1, 2, \ldots, M-1; \quad n = 1, 2, \ldots, M-2, \\ \dfrac{u_{m-1,2}-2u_{m2}+u_{m+1,2}}{h^2}+\dfrac{u_{m2}-2u_{m1}+u_{m0}}{h^2} = \varphi(mh, h), \\ \qquad\qquad m = 1, 2, \ldots, M-1, \\ u_{mn} = \psi(mh, nh), \qquad (x, y) \in \Gamma \end{cases} \tag{14}$$

paraît faire l'affaire. En effet, l'approximation a évidemment lieu du moment que le schéma (14) s'obtient en remplaçant les dérivées par des rapports aux différences, et la condition aux limites est approchée exactement. Chaque équation du premier groupe relie les valeurs de la solution en cinq points du réseau de la figure 19,*b*. Le second groupe relie pour m fixe les valeurs de la solution en cinq points de la figure 19,*c*.

Considérons dans leur ensemble les équations du premier groupe correspondant à n fixe, à savoir $n = 1$, et tout le second groupe. Le système ainsi formé est relatif aux valeurs u_{m1}, u_{m2} et u_{m0}, et u_{m0}, u_{01}, u_{02}, u_{M1}, u_{M2} sont définis par les conditions aux limites. Il se résout en définissant u_{m1}, u_{m2}, $m = 1, 2, \ldots, M-1$. Il faut ensuite utiliser l'équation aux différences du premier groupe correspondant à $n = 2$ et définir u_{m3} par une formule explicite en résolvant cette équation par rapport à l'unique inconnue u_{m3} qu'elle contient. En passant progressivement de u_{mn} à $u_{m,n+1}$ nous trouvons, en vertu des équations du premier groupe, les valeurs de $u^{(h)}$ en tous les points intérieurs du réseau. Quant aux valeurs aux points frontières, elles sont données dès le début.

Or, malgré tous les avantages qu'il offre, le schéma (14) est impraticable. Il est classique que la solution du problème de Dirichlet pour l'équation de Laplace est fonction en chaque point des valeurs de $\psi(x, y)|_\Gamma$ partout sur la frontière. Mais dans le cas du schéma aux différences construit, les valeurs de la solution $u^{(h)}$ en tous les points intérieurs se calculent sans avoir recours aux valeurs de $\psi(x, y)$ sur le côté supérieur du carré. Ce schéma aux différences est donc nécessairement divergent. Nous voyons que le caractère compliqué du schéma (13) découle du problème même.

En guise de conclusion soulignons une fois de plus que la condition de Courant, Friedrichs et Lewy ne suffit pas pour avoir la stabilité. Nous montrerons en particulier au §25 l'instabilité du schéma aux différences

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^n - u_{m-1}^n}{2h} = \varphi(mh, n\tau), \\ u_m^0 = \psi(mh), \end{cases}$$

quel que soit $r = \tau/h = \text{const}$. Ce schéma approche le problème de Cauchy

$$u_t - u_x = \varphi(x, t)$$
$$u(x, 0) = \psi(x),$$

pour lequel nous avons déjà considéré plusieurs autres schémas. D'autre part, on vérifie aisément que, pour $r \leqslant 1$, le schéma en question satisfait à la condition nécessaire de stabilité.

Prenons à cette fin, pour fixer les idées, le point (0, 1) du plan Oxt et supposons qu'il appartient au réseau D_h pour tous les h, de sorte que $1 = N\tau$, N entier. La valeur u_0^N se calcule par l'intermédiaire de u_{-1}^{N-1}, u_0^{N-1}, u_1^{N-1} qui s'obtiennent en fonction des cinq valeurs au niveau précédent $t = (N-2)\tau$, et ainsi de suite. On obtient finalement u_0^N par l'intermédiaire des valeurs $u_m^0 = \psi(mh)$, $m = -N, -N+1, \ldots, -1, 0, 1, \ldots, N$, aux points communs au réseau et au segment $-1/r \leqslant x \leqslant 1/r$ de l'axe Ox. Si $r = \tau/h < 1$, ce segment contient le point $x = 1$ en lequel la solution prend une valeur définissant $u(0, 1)$, $u(0, 1) = \psi(1)$. La condition de Courant, Friedrichs et Lewy est donc remplie pour $r \leqslant 1$.

Exercices

1. La solution de l'équation de la chaleur $u_t = u_{xx}$, $-\infty < x < \infty$, $t > 0$, est de la forme

$$u(x, t) = \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{t}} u(\xi, 0) e^{-\frac{(x-\xi)^2}{4t}} d\xi.$$

Existe-t-il un schéma aux différences convergent qui approche ce problème et est de la forme

$$\frac{u_m^{p+1} - u_m^p}{\tau} = \frac{1}{h^2} (\alpha_{-2} u_{m-2}^p + \alpha_{-1} u_{m-1}^p + \alpha_0 u_m^p + \alpha_1 u_{m+1}^p + \alpha_2 u_{m+2}^p),$$

α_j étant des constantes, si $\tau = h$?

2. Le système d'équations d'acoustique

$$\left.\begin{aligned} &\frac{\partial v}{\partial t} + \frac{\partial w}{\partial x} = 0, \\ &\frac{\partial w}{\partial t} + \frac{\partial v}{\partial x} = 0, \\ &v(x, 0) = \varphi(x), \quad w(x, 0) = \psi(x) \end{aligned}\right\} \quad t > 0, \quad -\infty < x < \infty,$$

admet une solution de la forme

$$\left.\begin{aligned} v(x, t) &= \frac{\varphi(x-t) + \psi(x-t) + \varphi(x+t) - \psi(x+t)}{2}, \\ w(x, t) &= \frac{\varphi(x-t) + \psi(x-t) - \varphi(x+t) + \psi(x+t)}{2}. \end{aligned}\right\}$$

Un schéma aux différences de la forme

$$\left.\begin{aligned} &\frac{v_m^{p+1} - v_m^p}{\tau} + \frac{w_{m+1}^p - w_m^p}{h} = 0, \quad p \geq 0, \\ &\frac{w_m^{p+1} - w_m^p}{\tau} + \frac{v_{m+1}^p - v_m^p}{h} = 0, \quad m = 0, \pm 1, \ldots, \\ &v_m^0 = \varphi(x_m), \quad w_m^0 = \psi(x_m) \end{aligned}\right\}$$

peut-il être convergent ? Comparer les domaines d'influence des données initiales du problème aux différences et de celui différentiel.

3. Le problème de Cauchy

$$\begin{aligned} &\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad t > 0, \quad -\infty < x < \infty, \\ &u(x, 0) = e^{i\alpha x}, \quad -\infty < x < \infty, \end{aligned}$$

a comme solution

$$u(x, t) = e^{i\alpha t} e^{i\alpha x}.$$

Le schéma aux différences correspondant

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad p = 0, 1, \ldots, \\ &u_m^0 = e^{i\alpha h m}, \quad m = 0, \pm 1, \ldots, \end{aligned}\right\}$$

admet pour solution

$$u_m^p = [1 - r + r e^{i\alpha h}]^p e^{i\alpha h m}$$

qui, quand $p = t/\tau$, $m = x/h$, tend vers la solution du problème différentiel pour h tendant vers zéro, quel que soit $r = \tau/h$ fixe. Or, pour $r > 1$, le schéma ne remplit pas la condition nécessaire de convergence de Courant, Friedrichs et Lewy. Expliquer ce paradoxe.

CHAPITRE 8

CERTAINS GRANDS PROCÉDÉS D'ÉTUDE DE LA STABILITÉ

§25. Analyse spectrale du problème de Cauchy aux différences

Nous allons exposer le très connu procédé d'étude des problèmes aux différences avec données initiales que nous devons à von Neumann. Le présent paragraphe sera consacré au problème de Cauchy aux différences à coefficients constants et certains résultats obtenus seront étendus au §26 au problème à coefficients variables.

1. Stabilité relativement aux données initiales. Nous avons déjà eu l'occasion de rencontrer à plusieurs reprises le problème

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1}-u_m^p}{\tau} - \dfrac{u_{m+1}^p - u_m^p}{h} = \varphi_m^p, & p = 0, 1, \ldots, [T/\tau]-1, \\ u_m^0 = \psi_m, & m = 0, \pm 1, \ldots \end{cases} \tag{1}$$

qui constitue l'exemple le plus simple du problème de Cauchy aux différences. En posant

$$f^{(h)} = \begin{cases} \varphi_m^p, & p = 0, 1, \ldots, [T/\tau]-1, \\ \psi_m, & m = 0, \pm 1, \ldots, \end{cases}$$

mettons (1) sous la forme

$$L_h u^{(h)} = f^{(h)}. \tag{2}$$

Définissons les normes $\|u^{(h)}\|_{U_h}$ et $\|f^{(h)}\|_{F_h}$ par

$$\|u^{(h)}\|_{U_h} = \max_p \max_m |u_m^p|, \quad \|f^{(h)}\|_{F_h} = \max_m |\psi_m| + \max_{m,p} |\varphi_m^p|.$$

La condition de stabilité du problème (2)

$$\|u^{(h)}\|_{U_h} \leqslant c\, \|f^{(h)}\|_{F_h} \tag{3}$$

s'écrit alors

$$\max_m |u_m^p| \leqslant c\left[\max_m |\psi_m| + \max_{m,k} |\varphi_m^k|\right], \quad p = 0, 1, \ldots, [T/\tau], \tag{4}$$

c étant indépendant de h (et de $\tau = rh$, $r = \text{const}$). On a la condition (4) pour $\{\psi_m\}$ et $\{\varphi_m^p\}$ quelconques. En particulier, pour que la stabilité ait lieu,

il faut que (4) soit remplie pour $\{\psi_m\}$ arbitraires, et $\varphi_m^p \equiv 0$ c.-à-d. que la solution du problème

$$\left.\begin{aligned} \frac{u_m^{p+1}-u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} &= 0, \quad p = 0, 1, \ldots, [T/\tau]-1, \\ u_m^0 &= \psi_m, \quad m = 0, \pm 1, \ldots, \end{aligned}\right\} \tag{5}$$

vérifie la condition

$$\max_m |u_m^p| \leqslant c \max_m |u_m^0|, \quad p = 0, 1, \ldots, [T/\tau], \tag{6}$$

$u_m^0 = \psi_m$ étant une fonction bornée quelconque.

La propriété (6) nécessaire pour qu'il y ait stabilité (4) du problème (1) s'appelle *stabilité du problème* (1) *par rapport à la perturbation des données initiales*. Elle signifie que la perturbation $\{u_m^0\}$ des données initiales de (1) induira une perturbation $\{u_m^p\}$ de la solution de ce problème, perturbation qui est, en vertu de (6), au plus c fois celle des données initiales, c étant indépendant de h.

2. Condition spectrale nécessaire de stabilité. Pour que le problème de Cauchy (1) soit stable par rapport aux données initiales, il faut que soit remplie la condition (6), en particulier si $\{u_m^0\}$ est un harmonique,

$$u_m^0 = e^{i\alpha m}, \quad m = 0, \pm 1, \ldots, \tag{7}$$

avec α un paramètre réel. Mais la solution du problème (5) avec la condition initiale (7) est de la forme

$$u_m^p = \lambda^p e^{i\alpha m}, \tag{8}$$

$\lambda = \lambda(\alpha)$ étant défini par substitution de (8) dans l'équation aux différences homogène du problème (5) :

$$\lambda(\alpha) = 1 - r + re^{i\alpha}, \quad r = \frac{\tau}{h} = \text{const.} \tag{9}$$

La solution (8) vérifie l'égalité

$$\max_m |u_m^p| = |\lambda(\alpha)|^p \max_m |u_m^0|.$$

Pour avoir la condition (6) il faut donc que, quels que soient α réels, on ait

$$|\lambda(\alpha)|^p \leqslant c, \quad p = 0, 1, \ldots, [T/\tau],$$

ou

$$|\lambda(\alpha)| \leqslant 1 + c_1\tau, \tag{10}$$

c_1 étant une constante indépendante de α et τ. C'est la *condition spectrale nécessaire de von Neumann* pour l'exemple considéré. Pourquoi « spectrale » ? L'existence d'une solution du type (8) montre que l'harmonique $\{e^{i\alpha m}\}$ est fonction propre de l'opérateur

$$u_m^{p+1} = (1-r)u_m^p + ru_{m+1}^p, \quad m = 0, \pm 1, \ldots,$$

qui fait correspondre, en vertu de (5), à la fonction $\{u_m^p\}$, $m = 0, \pm 1, \ldots$, définie au niveau $t_p = p\tau$ du réseau la fonction $\{u_m^{p+1}\}$, $m = 0, \pm 1, \ldots$,

définie au niveau $t_{p+1} = (p+1)\tau$. Le nombre $\lambda(\alpha) = 1-r+re^{i\alpha}$ est une valeur propre associée à l'harmonique $\{e^{i\alpha m}\}$ de l'opérateur u_m^{p+1}. La ligne parcourue par le point $\lambda(\alpha)$ dans le plan complexe, lorsque α parcourt l'axe réel, est formée de valeurs propres et constitue le spectre de l'opérateur en question.

Ainsi, la *condition nécessaire de stabilité* (10) peut s'énoncer comme suit : le spectre de l'opérateur de passage correspondant à l'équation aux différences du problème (5) doit se trouver dans un cercle de rayon $1+c_1\tau$ du plan complexe. Dans notre exemple, le spectre (9) ne dépend pas de τ. Aussi la condition (10) équivaut-elle à l'exigence que le spectre $\lambda(\alpha)$ se situe dans le cercle unité

$$|\lambda(\alpha)| \leqslant 1. \tag{11}$$

Analysons la stabilité du problème (1) à la lumière du critère formulé. Le spectre (9) est une circonférence de centre au point $1-r$ et de rayon r du plan complexe. Si $r < 1$, cette circonférence est intérieure au cercle unité (leur point de contact est $\lambda = 1$), si $r = 1$, elle se confond avec la circonférence unité et pour $r > 1$ elle est extérieure au cercle unité (fig. 20).

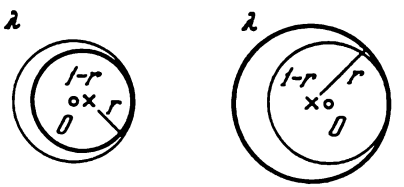


Fig. 20

La condition nécessaire de stabilité (11) est donc remplie pour $r \leqslant 1$ et elle l'est plus lorsque $r > 1$. Au n° 3 du §21 nous avons déjà examiné le schéma aux différences proposé et démontré sa stabilité pour $r \leqslant 1$ et son instabilité pour $r > 1$. Ainsi, la condition nécessaire de stabilité de von Neumann s'est trouvée dans ce cas suffisamment sensible pour séparer nettement le cas stable de celui instable.

Dans le cas général du problème de Cauchy pour les équations et systèmes aux différences le critère spectral nécessaire de stabilité de von Neumann est que le spectre $\lambda = \lambda(\alpha, h)$ du problème aux différences se trouve, pour tous les h suffisamment petits, dans le cercle

$$|\lambda| \leqslant 1+\varepsilon \tag{12}$$

du plan complexe, le nombre positif ε choisi à l'avance étant aussi petit que l'on veut.

Notons que si, pour un problème aux différences, le spectre s'avère indépendant de h (et de τ), la condition (12) équivaut à exiger que le spectre $\lambda = \lambda(\alpha, h) = \lambda(\alpha)$ se trouve dans le cercle unité

$$|\lambda(\alpha)| \leqslant 1. \tag{12'}$$

Le spectre d'un problème aux différences avec la condition (12) est assimilé à l'ensemble de tous les $\lambda = \lambda(\alpha, h)$ tels que l'équation (ou le système) aux différences homogène correspondant admette une solution de la forme

$$u_m^p = [\lambda(\alpha, h)]^p [u^0 e^{i\alpha m}], \quad m = 0, \pm 1, \ldots, \tag{13}$$

où u^0 est un nombre (l'unité) s'il s'agit d'une équation scalaire, et un vecteur numérique dans le cas d'une équation vectorielle, i.e. d'un système d'équations aux différences scalaires.

Si la condition nécessaire de von Neumann (12) n'a pas lieu, le choix des normes a beau être raisonnable : on n'aura jamais la stabilité. Si cette condition est par contre remplie, on peut s'attendre à la stabilité pour un choix raisonnable des normes. Une question analogue (l'indépendance du critère spectral de stabilité par rapport au choix des normes) a déjà été le sujet d'une discussion dans le cas de schémas aux différences pour les équations différentielles ordinaires (§15).

3. Exemples. Considérons quelques problèmes de Cauchy aux différences qui présentent de l'intérêt et dont nous analyserons la stabilité au travers du critère spectral de von Neumann. Commençons par des schémas approchant le problème de Cauchy différentiel

$$\left.\begin{aligned} \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} &= \varphi(x, t), \quad -\infty < x < \infty, \ 0 < t < T, \\ u(x, 0) &= \psi(x), \quad -\infty < x < \infty. \end{aligned}\right\} \tag{14}$$

Exemple 1. Soit le schéma aux différences

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_m^p - u_{m-1}^p}{h} = \varphi(x_m, t_p), \quad p = 0, 1, \ldots, [T/\tau]-1,$$

$$u_m^0 = \psi(x_m), \quad m = 0, \pm 1, \ldots$$

Portons une expression de la forme (8) dans l'équation aux différences homogène correspondante, on obtient par de simples transformations

$$\lambda(\alpha) = 1 + r - re^{-i\alpha}.$$

Le spectre est donc une circonférence centrée en $1+r$ et de rayon r

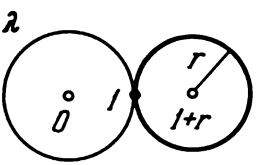


Fig. 21

(fig. 21). Pour aucun r le spectre n'appartient au cercle unité. On n'a jamais la condition de stabilité (12').

Nous avons établi (§24) qu'aucun r ne vérifie la condition nécessaire de convergence (et de stabilité) de Courant, Friedrichs et Lewy.

EXEMPLE 2. Considérons le schéma aux différences

$$\left.\begin{aligned} \frac{u_m^{p+1}-u_m^p}{\tau}-\frac{u_{m+1}^p-u_{m-1}^p}{2h}-\frac{\tau}{2h^2}(u_{m-1}^p-2u_m^p+u_{m+1}^p) &= \varphi(x_m, t_p), \\ u_m^0 &= \psi_m \end{aligned}\right\} \quad (15)$$

approchant le problème (14) à l'ordre deux en h (§22). $\lambda = \lambda(\alpha)$ est alors défini à partir de l'équation

$$\frac{\lambda-1}{\tau}-\frac{e^{i\alpha}-e^{-i\alpha}}{2h}-\frac{\tau}{2h^2}(e^{i\alpha}-2+e^{-i\alpha}) = 0.$$

Notons toujours $r = \tau/h$. Vu que

$$\frac{e^{i\alpha}-e^{-i\alpha}}{2i} = \sin\alpha,$$

$$\frac{e^{i\alpha}-2+e^{-i\alpha}}{4} = -\left(\frac{e^{\frac{i\alpha}{2}}-e^{-\frac{i\alpha}{2}}}{2i}\right)^2 = -\sin^2\frac{\alpha}{2},$$

nous obtenons

$$\lambda = 1+ir\sin\alpha-2r^2\sin^2\frac{\alpha}{2}, \quad (16)$$

$$|\lambda(\alpha)|^2 = \left(1-2r^2\sin^2\frac{\alpha}{2}\right)^2+r^2\sin^2\alpha,$$

et par de simples transformations

$$1-|\lambda|^2 = 4r^2\sin^4\frac{\alpha}{2}(1-r^2). \quad (17)$$

La condition de von Neumann est remplie si le second membre est non négatif, $r \leqslant 1$, et elle ne l'est pas pour $r > 1$.

EXEMPLE 3. Prenons pour le même problème de Cauchy (14) un autre schéma aux différences, à savoir

$$L_h u^{(h)} = \left\{\begin{aligned} \frac{u_m^{p+1}-u_m^p}{\tau}-\frac{u_{m+1}^p-u_{m-1}^p}{2h} &= \varphi(x_m, t_p), \\ u_m^0 &= \psi(x_m). \end{aligned}\right. \quad (18)$$

Portons (8) dans l'équation (18) et simplifions ; cela donne une équation pour λ :

$$\frac{\lambda-1}{\tau}-\frac{e^{i\alpha}-e^{-i\alpha}}{2h} = 0$$

ou

$$\lambda(\alpha) = 1+i\left(\frac{\tau}{h}\sin\alpha\right).$$

Le spectre $\lambda = \lambda(\alpha)$ occupe le segment vertical de longueur $2\tau/h$ passant par le point $\lambda = 1$ (fig. 22).

Si $\tau/h = r =$ const, la condition (12′) n'a pas lieu (le spectre ne se trouve pas dans le cercle unité). Si, pour $h \to 0$, le pas τ varie comme h^2 de sorte

que $\tau = rh^2$, alors le point $\lambda(\alpha)$ le plus éloigné de $\lambda = 0$ est de module

$$|\lambda(\alpha)|_{\alpha=\pi/2} = \sqrt{1+\left(\frac{\tau}{h}\right)^2} = \sqrt{1+\tau r} \leqslant 1+\frac{r}{2}\tau.$$

Dans ce cas, la condition $|\lambda(\alpha)| \leqslant 1+c\tau$ est satisfaite pour $c = r/2$.

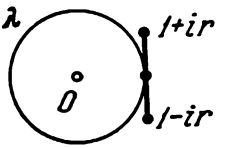


Fig. 22

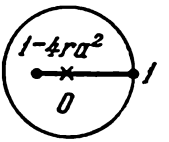


Fig. 23

La contrainte $\tau = rh^2$ sur la décroissance du pas en temps τ pour $h \to 0$ est sans doute plus sévère que $\tau \doteq rh$, $r \leqslant 1$, condition sous laquelle on a le critère de von Neumann pour les schémas aux différences (5) et (15) approchant le problème de Cauchy (14).

A la fin du §24 nous avons montré que le critère de Courant, Friedrichs et Lewy ne permet de conclure à l'instabilité que pour $\tau/h > 1$ et qu'il n'affirme rien lorsque $\tau/h \leqslant 1$. Ce critère est donc plus faible que celui de von Neumann.

Revenons maintenant à deux schémas aux différences construits au §22 qui approchent le problème de Cauchy relatif à l'équation de la chaleur

$$\left.\begin{aligned} u_t - a^2 u_{xx} &= \varphi(x, t), \quad -\infty < x < \infty, \ 0 < t < T, \\ u(x, 0) &= \psi(x), \quad -\infty < x < \infty. \end{aligned}\right\} \tag{19}$$

Exemple 4. En portant $u_m^p = \lambda^p e^{i\alpha m}$ dans l'équation aux différences homogène correspondante, le schéma explicite

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - a^2 \dfrac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \varphi(mh, n\tau), \\ u_m^0 = \psi(mh), \quad m = 0, \pm 1, \ldots; \quad p = 0, 1, \ldots, [T/\tau]-1, \end{cases}$$

conduit à la relation

$$\frac{\lambda - 1}{\tau} - a^2 \frac{e^{-i\alpha} - 2 + e^{i\alpha}}{h^2} = 0.$$

Vu que

$$\frac{e^{-i\alpha} - 2 + e^{i\alpha}}{4} = -\left(\frac{e^{\frac{i\alpha}{2}} - e^{-\frac{i\alpha}{2}}}{2i}\right)^2 = -\sin^2\frac{\alpha}{2},$$

nous obtenons

$$\lambda(\alpha) = 1 - 4ra^2 \sin^2\frac{\alpha}{2}, \quad r = \frac{\tau}{h^2}.$$

Lorsque α varie, le nombre $\lambda(\alpha)$ parcourt le segment $1 - 4ra^2 \leqslant \lambda \leqslant 1$ de l'axe réel (fig. 23).

La stabilité exige que l'extrémité gauche de ce segment se situe dans le cercle unité $1-4ra^2 \geqslant -1$ ou

$$r \leqslant \frac{1}{2a^2}. \tag{20}$$

Si $r > \frac{1}{2a^2}$, le point $\lambda(\alpha) = 1-4ra^2 \sin^2 \frac{\alpha}{2}$ correspondant à $\alpha = \pi$ est à gauche de -1. L'harmonique $e^{inm} = (-1)^m$ engendre la solution

$$u_m^p = (1-4a^2r)^p (-1)^m$$

qui ne vérifie (6) pour aucune constante c.

EXEMPLE 5. Soit le schéma

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1}-u_m^p}{\tau} - a^2 \dfrac{u_{m+1}^{p+1}-2u_m^{p+1}+u_{m-1}^{p+1}}{h^2} = \varphi(mh, n\tau), \\ \qquad\qquad\qquad\qquad\qquad\qquad u_m^0 = \psi(mh), \\ m = 0, \pm 1, \ldots ; \quad p = 0, 1, \ldots, [T/\tau]-1. \end{cases} \tag{21}$$

Des calculs analogues aboutissent à l'expression

$$\lambda(\alpha) = \frac{1}{1+4ra^2 \sin^2 \frac{\alpha}{2}}, \quad r = \frac{\tau}{h^2}. \tag{22}$$

Le spectre de ce problème remplit le segment

$$\left[1+4ra^2 \sin^2 \frac{\alpha}{2}\right]^{-1} \leqslant \lambda \leqslant 1$$

de l'axe réel, et la condition $|\lambda| \leqslant 1$ est satisfaite pour tout r.

Le critère spectral de von Neumann reste valable dans le cas de plusieurs variables d'espace.

EXEMPLE 6. Pour le problème

$$\left.\begin{aligned} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad t > 0, \\ u(x, y, t) = \psi(x, y) \end{aligned}\right\}$$

prenons le réseau $(x_m, y_n, t_p) = (mh, nh, p\tau)$. Remplaçons les dérivées par des rapports aux différences, il vient le schéma

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{mn}^{p+1}-u_{mn}^p}{\tau} - \dfrac{u_{m+1,n}^p-2u_{mn}^p+u_{m-1,n}^p}{h^2} - \dfrac{u_{m,n+1}^p-2u_{mn}^p+u_{m,n-1}^p}{h^2} = 0, \\ u_{mn}^0 = \psi(mh, nh). \end{cases} \tag{23}$$

Définissant $u_{mn}^0 = e^{i(\alpha m+\beta n)}$, i.e. sous forme d'un harmonique bidimensionnel dépendant de deux paramètres réels α et β, on trouve une solution de la forme

$$u_{mn}^p = \lambda^p(\alpha, \beta) e^{i(\alpha m+\beta n)}.$$

En substituant cette expression dans l'équation aux différences, en simpli-

fiant et en effectuant des transformations identiques, on trouve

$$\lambda(\alpha, \beta) = 1 - 4r \sin^2 \frac{\alpha}{2} - 4r \sin^2 \frac{\beta}{2}.$$

Avec α et β variant, le point $\lambda = \lambda(\alpha, \beta)$ parcourt le segment

$$1 - 8r \leqslant \lambda \leqslant 1$$

de l'axe réel. La condition de stabilité est remplie si $1 - 8r \geqslant -1$, $r \leqslant {}^1/_4$.

Citons un exemple illustrant l'application du critère de von Neumann à des équations aux différences relatives aux valeurs de la fonction cherchée non pas à deux mais à trois niveaux temporels.

EXEMPLE 7. Le problème de Cauchy pour l'équation des ondes

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \quad -\infty < x < \infty, \quad 0 < t < T,$$

$$u(x, 0) = \psi_1(x), \quad \frac{\partial u(x, 0)}{\partial t} = \psi_2(x), \quad -\infty < x < \infty,$$

est approché par le schéma aux différences

$$\left.\begin{aligned} &\frac{u_m^{p+1} - 2u_m^p + u_m^{p-1}}{\tau^2} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\ &p = 1, 2, \ldots, [T/\tau] - 1, \\ &u_m^0 = \psi_1(x_m), \quad \frac{u_m^1 - u_m^0}{\tau} = \psi_2(x_m), \quad m = 0, \pm 1, \ldots \end{aligned}\right\} \tag{24}$$

Portons dans l'équation aux différences une solution de la forme (8). Après de simples transformations nous arrivons à l'équation suivante définissant λ :

$$\lambda^2 - 2\left(1 - 2r^2 \sin^2 \frac{\alpha}{2}\right)\lambda + 1 = 0, \quad r = \frac{\tau}{h}.$$

Le produit des racines de cette équation vaut 1. Si le discriminant

$$d(\alpha) = 4r^2 \sin^2 \alpha \left(r^2 \sin^2 \frac{\alpha}{2} - 1\right)$$

de l'équation du second degré est négatif, les racines $\lambda_1(\alpha)$ et $\lambda_2(\alpha)$ sont complexes conjuguées et de module égal à l'unité. Pour $r < 1$, le discriminant reste négatif quel que soit α. La figure 24, *a* représente le spectre dans ce cas, qui remplit une partie de la circonférence unité. Si $r = 1$, le spectre occupe toute la circonférence et si $r > 1$, avec α variant de 0 à π, les racines $\lambda_1(\alpha)$ et $\lambda_2(\alpha)$ se déplacent à partir du point $\lambda = 1$ suivant la circonférence unité (l'une dans le sens des aiguilles d'une montre et l'autre dans le sens contraire) jusqu'à ce qu'elles se confondent en $\lambda = -1$, après quoi l'une des racines suit l'axe réel à gauche et l'autre à droite puisqu'elles sont réelles et $\lambda_1\lambda_2 = 1$ (fig. 24,*b*). La condition de stabilité est remplie pour $r \leqslant 1$.

Considérons le problème de Cauchy pour le système hyperbolique

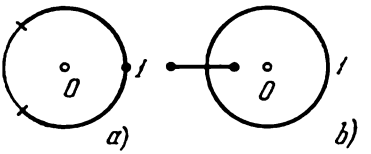


Fig. 24

d'équations différentielles décrivant la propagation du son

$$\left.\begin{aligned} &\frac{\partial v}{\partial t} = \frac{\partial w}{\partial x}, \\ &\qquad\qquad -\infty < x < \infty, \quad 0 < t < T, \\ &\frac{\partial w}{\partial t} = \frac{\partial v}{\partial x}, \\ &v(x, 0) = \psi_1(x), \quad w(x, 0) = \psi_2(x), \quad -\infty < x < \infty. \end{aligned}\right\} \tag{25}$$

Posons

$$u(x, t) = \begin{pmatrix} v(x, t) \\ w(x, t) \end{pmatrix}, \quad \psi(x) = \begin{pmatrix} \psi_1(x) \\ \psi_2(x) \end{pmatrix}$$

et écrivons (25) sous forme vectorielle :

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} - A\frac{\partial u}{\partial x} = 0, \qquad -\infty < x < \infty, \quad 0 < t < T, \\ &u(x, 0) = \psi(x), \quad -\infty < x < \infty, \end{aligned}\right\} \tag{25'}$$

où

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Examinons deux schémas aux différences approchant le problème (25').

Exemple 8. Soit le schéma aux différences

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - A\,\frac{u_{m+1}^p - u_m^p}{h} = 0, \qquad p = 0, 1, \ldots, [T/\tau]-1, \\ &\qquad u_m^0 = \psi(x_m), \quad m = 0, \pm 1, \ldots \end{aligned}\right\} \tag{26}$$

Cherchons la solution de l'équation aux différences homogène vectorielle sous la forme (13) :

$$u_m^p = \lambda^p(u^0 e^{i\alpha m}) = \lambda^p \begin{pmatrix} v^0 \\ w^0 \end{pmatrix} e^{i\alpha m}$$

et portons cette expression dans l'équation aux différences (26). Il vient l'égalité

$$\frac{\lambda - 1}{\tau} u^0 - A\,\frac{e^{i\alpha} - 1}{h} u^0 = 0$$

ou

$$(\lambda-1)u^0-r(e^{i\alpha}-1)Au^0 = 0, \quad r = \frac{\tau}{h}, \tag{27}$$

qu'on peut prendre pour écriture vectorielle d'un système d'équations linéaires servant à déterminer les composantes du vecteur u^0.

Ecrivons le système (27) sous forme développée

$$\begin{pmatrix} \lambda-1 & -r(e^{i\alpha}-1) \\ -r(e^{i\alpha}-1) & \lambda-1 \end{pmatrix} \begin{pmatrix} v^0 \\ w^0 \end{pmatrix} = 0. \tag{28}$$

Le système d'équations linéaires (28) n'admet une solution non triviale $u^0 = \begin{pmatrix} v^0 \\ w^0 \end{pmatrix}$ que pour des $\lambda = \lambda(\alpha)$ tels que le déterminant de (28) s'annule :

$$(\lambda-1)^2 = r^2(e^{i\alpha}-1)^2.$$

D'où

$$\lambda_1(\alpha) = 1-r+re^{i\alpha},$$
$$\lambda_2(\alpha) = 1+r-re^{i\alpha}.$$

Les racines $\lambda_1(\alpha)$ et $\lambda_2(\alpha)$ parcourent les circonférences de rayon r et de centres $1-r$ et $1+r$ respectivement (fig. 25). Aucun r ne vérifie la condition de stabilité de von Neumann.

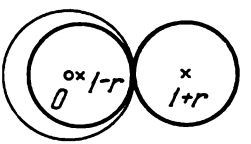


Fig. 25

Exemple 9. Considérons le schéma aux différences

$$\left.\begin{aligned} &\frac{u_m^{p+1}-u_m^p}{\tau} - A\,\frac{u_{m+1}^p-u_{m-1}^p}{2h} - \frac{\tau}{2h^2}\,A^2(u_{m+1}^p-2u_m^p+u_{m-1}^p) = 0, \\ &p = 0, 1, \ldots, [T/\tau]-1\ ; \qquad m = 0, \pm 1, \ldots, \\ &u_m^0 = \psi(x_m), \qquad\qquad m = 0, \pm 1, \ldots, \end{aligned}\right\} \tag{29}$$

qui approche le problème (25′) à l'ordre 2 et est analogue au schéma (15) pour le cas scalaire (14). La condition pour que l'équation vectorielle (25) admette une solution non triviale de la forme (13) est que, comme dans l'exemple 8, le déterminant du système pour définir $u^0 = \begin{pmatrix} v^0 \\ w^0 \end{pmatrix}$ devienne nul. En l'égalant à zéro on obtient une équation du second degré en $\lambda = \lambda(\alpha)$, d'où l'on tire

$$\left.\begin{aligned} \lambda_1 &= 1+ir\sin\alpha-2r^2\sin^2\frac{\alpha}{2}, \\ \lambda_2 &= 1-ir\sin\alpha-2r^2\sin^2\frac{\alpha}{2}. \end{aligned}\right\} \tag{30}$$

Ces formules étant analogues à (16), on obtient à la place de (17)

$$1-|\lambda_{1,2}(\alpha)|^2 = 4r^2\sin^4\frac{\alpha}{2}(1-r^2).$$

Le spectre donné par les formules (30) se trouve dans le cercle unité pour $r < 1$.

4. Représentation intégrale d'une solution (*). Abordons un problème de Cauchy de la forme

$$\left.\begin{aligned} b_{-1}u_{m-1}^{p+1}+b_0u_m^{p+1}+b_1u_{m+1}^{p+1}-(a_{-1}u_{m-1}^p+a_0u_m^p+a_1u_{m+1}^p) &= \tau\varphi_m^p, \\ p = 0, 1, \ldots, [T/\tau]-1, \\ u_m^0 = \psi_m, \quad m = 0, \pm 1, \ldots, \end{aligned}\right\} \tag{31}$$

à coefficients constants en supposant que

$$b_{-1}e^{-i\alpha}+b_0+b_1e^{i\alpha} \neq 0, \quad 0 \leq \alpha \leq 2\pi. \tag{32}$$

Les schémas aux différences (1), (15), (18), (19), (21) se ramènent à la forme (31) en multipliant par τ les deux membres des équations aux différences y figurant.

Notons avant tout que pour des fonctions discrètes bornées $\{\varphi_m^p\}$ et $\{\psi_m\}$ quelconques, le problème (31) admet une solution bornée et une seule. En effet, connaissant que, pour p donné fixe, $\{u_m^p\}$ existe et est borné, l'équation (31) devient l'équation aux différences ordinaire du second ordre

$$b_{-1}u_{m-1}^{p+1}+b_0u_m^{p+1}+b_1u_{m+1}^{p+1} = \varphi_m^p+(a_{-1}u_{m-1}^p+a_0u_m^p+a_1u_{m+1}^p) \tag{33}$$

en $\{u_m^{p+1}\}$ avec le second membre borné. L'équation caractéristique correspondante $b_{-1}+b_0q+b_1q^2 = 0$ ne possède pas, selon (32), de racines $q = e^{i\alpha}$ de module égal à l'unité. Aussi admet-elle, comme on l'a montré fin du n° 2 §3, une seule solution bornée $\{u_m^{p+1}\}$. Or, $\{u_m^0\} = \{\psi_m\}$ est défini et borné et on définit donc successivement de façon unique, à partir de (33), les fonctions bornées $\{u_m^1\}$, $\{u_m^2\}$, ... Nous aurons besoin des renseignements suivants sur les séries de Fourier.

A chaque suite de nombres c_m, $m = 0, \pm 1, \ldots$, telle que $\sum_m |c_m|^2 < \infty$, correspond une série de Fourier convergente (en moyenne quadratique)

$$\frac{1}{\sqrt{2\pi}}\sum_m c_m e^{-i\alpha m} = C(\alpha), \tag{34}$$

dont la somme est la fonction $C(\alpha)$, $\int_{-\infty}^{\infty}|C(\alpha)|^2\,d\alpha < \infty$, de carré intégrable sur le segment $0 \leq \alpha \leq 2\pi$. Inversement, toute fonction $C(\alpha)$ de carré intégrable sur le segment $0 \leq \alpha \leq 2\pi$ se développe de façon unique en une

(*) Les résultats de ce numéro ne sont pas utilisés dans les paragraphes suivants.

série de Fourier (34) à coefficients c_m obtenus par les formules

$$c_m = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} C(\alpha) e^{i\alpha m}\, d\alpha. \tag{35}$$

On a alors l'*égalité de Parseval*

$$\int_0^{2\pi} |C(\alpha)|^2\, d\alpha = \sum_m |c_m|^2. \tag{36}$$

THÉORÈME 1. *Soit, dans le problème* (31),

$$\max_p \sum_m |\varphi_m^p|^2 < \infty, \quad \sum_m |\psi_m|^2 < \infty.$$

La solution bornée de ce problème se met alors sous forme intégrale

$$u_m^p = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} U^p(\alpha) e^{i\alpha m}\, d\alpha, \tag{37}$$

où la fonction de carré intégrable $U^p(\alpha)$ est définie par la relation de récurrence

$$U^{p+1}(\alpha) = \lambda(\alpha)\, U^p(\alpha) + \frac{\tau}{b_{-1}e^{-i\alpha} + b_0 + b_1 e^{i\alpha}}\, \Phi^p(\alpha). \tag{38}$$

Ici

$$\Phi^p(\alpha) = \frac{1}{\sqrt{2\pi}} \sum_m \varphi_m^p e^{-i\alpha m},$$

$$U^0(\alpha) \equiv \Psi(\alpha) = \frac{1}{\sqrt{2\pi}} \sum_m \psi_m e^{-i\alpha m}$$

et la fonction

$$\lambda(\alpha) \equiv \frac{a_{-1}e^{-i\alpha} + a_0 + a_1 e^{i\alpha}}{b_{-1}e^{-i\alpha} + b_0 + b_1 e^{i\alpha}}$$

est choisie de façon que pour chaque α, $0 \leqslant \alpha \leqslant 2\pi$, la fonction discrète $v_m^p \equiv \lambda^p(\alpha) e^{i\alpha m}$ vérifie l'équation homogène correspondant à (31).

La démonstration se fait en substituant directement (37) dans l'équation et la condition initiale (31) moyennant les égalités (34) et (35).

CONSÉQUENCE. *Si dans* (31) *la fonction $\varphi_m^p \equiv 0$, alors $\Phi^p(\alpha) \equiv 0$; on a, en vertu de* (38), $U^p(\alpha) = \lambda^p(\alpha)\, \Psi(\alpha)$ *et* (37) *entraîne*

$$u^p(\alpha) = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \lambda^p(\alpha)\, \Psi(\alpha) e^{i\alpha m}\, d\alpha. \tag{39}$$

On utilise les représentations intégrales (37) et (39) pour analyser les propriétés du schéma aux différences (31).

Définissons les normes par les égalités

$$\|u^p\|^2 = \sum_m |u_m^p|^2; \quad \|u^{(h)}\|_{U_h} = \max \|u^p\|;$$

$$\|\varphi^p\|^2 = \sum_m |\varphi_m|^2; \qquad \|\psi\|^2 = \sum_m |\psi_m|^2; \tag{40}$$

$$\|f^{(h)}\|_{F_h} = \left\| \begin{matrix} \varphi_m^p \\ \psi_m \end{matrix} \right\|_{F_h} = \|\psi\| + \max_p \|\varphi^p\|.$$

THÉORÈME 2. *Pour que le schéma aux différences* (31) *soit stable par rapport aux données initiales, i.e. pour qu'on ait les inégalités*

$$\|u^p\| \leqslant c\,\|u^0\|, \quad p = 0, 1, \ldots, [T/\tau], \tag{40a}$$

avec $\{u_m^0\} = \{\psi_m\}$ *arbitraire,* $\|\psi\| < \infty$, $\varphi_m^p \equiv 0$ *et* c *une constante indépendante de* h *(et de* $\tau = \tau(h)$*), il faut et il suffit que le spectre* $\lambda = \lambda(\alpha)$ *se trouve dans le cercle* (10) :

$$|\lambda(\alpha)| \leqslant 1 + c_1\tau, \tag{41}$$

c_1 *ne dépendant pas de* h *(et de* τ*).*

DÉMONSTRATION. Etablissons d'abord la *suffisance*. Dans la condition (41) on a évidemment

$$|\lambda(\alpha)|^p \leqslant |1 + c_1\tau|^{T/\tau} e^{c_1 T}. \tag{42}$$

Vu l'égalité de Parseval et l'inégalité (42), la représentation (39) entraîne

$$\|u^p\| = \left[\int_0^{2\pi} |\lambda^p(\alpha)\, U(\alpha)|^2\, d\alpha\right]^{1/2} \leqslant$$

$$\leqslant e^{c_1 T}\left[\int_0^{2\pi} |U(\alpha)|^2\, d\alpha\right]^{1/2} = e^{c_1 T}\,\|u^0\| = c\,\|u^0\|.$$

Nécessité. Montrons que l'absence de (41) pour tout c_1 fixe entraîne l'instabilité. Bien que la solution

$$u_m^p = \lambda^p(\alpha) e^{i\alpha m}, \quad p = 0, \ldots, [T/\tau],$$

ne soit pas alors bornée pour $\tau \to 0$, on ne saurait utiliser cette propriété pour démontrer l'instabilité dans la norme choisie (40), parce que $\{e^{i\alpha m}\}$ n'appartient pas à l'espace de fonctions définies sur le réseau dont la somme des carrés des modules est bornée.

L'instabilité se démontre en remarquant en premier lieu qu'il existe toujours une fonction de carré intégrable $U(\alpha)$ telle que

$$\frac{1}{2\pi}\int_0^{2\pi} |\lambda(\alpha)|^{2p}\, |U(\alpha)|^2\, d\alpha \geqslant$$

$$\geqslant \max_\alpha [|\lambda(\alpha)|^{2p} - \varepsilon] \cdot \frac{1}{2\pi}\int_0^{2\pi} |U(\alpha)|^2\, d\alpha, \tag{43}$$

avec $\varepsilon > 0$ arbitraire. En effet, si $\max_\alpha |\lambda(\alpha)| = |\lambda(\alpha^*)|$, on pose

$$U(\alpha) = \begin{cases} 1 \text{ si } \alpha \in [\alpha^*-\delta, \alpha^*+\delta], \\ 0 \text{ si } \alpha \notin [\alpha^*-\delta, \alpha^*+\delta]. \end{cases}$$

Vu la continuité de $|\lambda(\alpha)|^{2p}$, pour $\delta = \delta(\varepsilon)$ suffisamment petit on a l'inégalité (43). Si (42) n'a pas lieu, il existe une suite h_k et une suite correspondante $\tau_k = \tau(h_k)$ telles que

$$c_k \equiv \left[\max_\alpha |\lambda(\alpha, h_k)|\right]^{[T/\tau_k]} \to \infty \quad \text{pour} \quad k \to \infty.$$

Posons $\varepsilon = 1$ et choisissons une $U(\alpha)$ vérifiant (43). Prenons pour suite $\{u_m^0\}$ la suite de coefficients de Fourier de $U(\alpha)$. Alors (43) s'écrit lorsque $p = [T/\tau]$

$$\|u^{[T/\tau]}\| \geqslant (c_k-1)\,\|u^0\|, \quad c_k \to \infty, \quad \text{pour} \quad h \to 0,$$

ce qui équivaut à l'absence de la stabilité (40a) par rapport aux données initiales.

THÉORÈME 3. *Le problème de Cauchy aux différences* (31) *est stable pour le choix* (40) *des normes si et seulement s'il vérifie le critère spectral de stabilité* (41).

DÉMONSTRATION. La *nécessité* est évidente parce que si ce critère n'est pas satisfait, il n'y a pas, en vertu du théorème 2, de stabilité par rapport aux données initiales.

Suffisance. Etablissons que tout $k \geqslant 0$ vérifie l'inégalité

$$\|u^{k+1}\| \leqslant (1+c_1\tau)\,\|u^k\| + c_2\tau \max_n \|\varphi^n\|, \tag{44}$$

d'où résultent, quel que soit j, $p \geqslant j \geqslant 0$, les inégalités

$$(1+c_1\tau)^j\,\|u^{p+1-j}\| \leqslant (1+c_1\tau)^{j+1}\,\|u^{p-j}\| + c_2\tau(1+c_1\tau)^p\,\|\varphi^p\|.$$

En sommant terme à terme en $j = 0, 1, \ldots, p$ et en rejetant les mêmes termes dans les deux membres on écrit

$$\|u^{p+1}\| \leqslant (1+c_1\tau)^{j+1}\,\|u^{p-j}\| + c_2\tau p(1+c_1\tau)^p \max \|\varphi^p\| \leqslant$$
$$\leqslant e^{c_1T}\,\|u^0\| + c_2Te^{c_1T} \max_n \|\varphi^n\| \leqslant \text{const}\,\|f^{(h)}\|_{F_h}.$$

D'où la stabilité du moment que p, $p = 0, 1, \ldots, [T/\tau]-1$, est arbitraire.

On démontre l'inégalité (44) en recourant à la représentation intégrale (37) de la solution et à la relation de récurrence (38), d'où

$$u_m^{k+1} = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} U^{k+1}(\alpha)e^{i\alpha m}\,d\alpha = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \lambda(\alpha)\,U^k(\alpha)e^{i\alpha m}\,d\alpha +$$
$$+ \frac{\tau}{\sqrt{2\pi}} \int_0^{2\pi} \frac{\Phi^p(\alpha)}{b_{-1}e^{-i\alpha}+b_0+b_1e^{i\alpha}}\,e^{i\alpha m}\,d\alpha. \tag{45}$$

Ainsi, la fonction discrète $\{u_m^{k+1}\}$ de l'argument m se décompose en la somme de deux fonctions discrètes écrites sous forme des intégrales par rapport au paramètre α.

En vertu de l'égalité de Parseval on écrit pour les normes de ces deux fonctions

$$\left\| \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \lambda(\alpha)\, U^k(\alpha) e^{i\alpha m}\, d\alpha \right\| = \left[\int_0^{2\pi} |\lambda(\alpha)\, U^k(\alpha)|^2\, d\alpha \right]^{1/2} \leqslant$$

$$\leqslant \max_\alpha |\lambda(\alpha)| \left[\int_0^{2\pi} |U^k(\alpha)|^2\, d\alpha \right]^{1/2} \leqslant (1 + c_1\tau)\, \|u^k\|;$$

$$\left\| \frac{\tau}{\sqrt{2\pi}} \int_0^{2\pi} \frac{\Phi^k(\alpha)}{b_{-1}e^{-i\alpha} + b_0 + b_1 e^{i\alpha}} e^{i\alpha m}\, d\alpha \right\| =$$

$$= \tau \left[\int_0^{2\pi} \left| \frac{\Phi^k(\alpha)}{b_{-1}e^{-i\alpha} + b_0 + b_1 e^{i\alpha}} \right|^2 d\alpha \right]^{1/2} \leqslant \frac{\tau}{\min |b_{-1}e^{-i\alpha} + b_0 + b_1 e^{i\alpha}|} \times$$

$$\times \left[\int_0^{2\pi} |\Phi^k(\alpha)|^2\, d\alpha \right]^{1/2} = \frac{\tau}{\min |b_{-1}e^{-i\alpha} + b_0 + b_1 e^{i\alpha}|} \|\varphi^k\| \leqslant \tau c_2 \max_k \|\varphi^k\|.$$

A partir de deux dernières estimations des normes des termes figurant dans le second membre de (45) on obtient l'estimation (44) qui achève la démonstration.

On montre qu'en prenant pour norme non pas (40), mais $\|u^p\| = \sup_m |u_m^p|$, le critère spectral $|\lambda(\alpha)| < 1 + c_1\tau$ ne suffit plus pour avoir la stabilité. C'est également vrai pour le problème de Cauchy aux différences dans le cas d'un système d'équations.

On se sert de la représentation intégrale (37) de la solution du problème de Cauchy aux différences non seulement pour établir la stabilité, mais aussi pour révéler d'autres propriétés du schéma aux différences.

Supposons par exemple que le spectre $\lambda = \lambda(\alpha)$ est pour $\alpha \neq 0$ strictement intérieur au cercle unité. Les solutions $u_m^p = \lambda^p(\alpha) e^{i\alpha m}$, $\alpha \neq 0$, s'amortissent alors par multiplication par $\lambda(\alpha)$ en passant d'un niveau à l'autre. La formule (39) montre que lorsque $[T/\tau] = p$, on obtient une fonction discrète associée à $\lambda^p(\alpha)\Psi(\alpha)$, qui se concentre sur des ondes longues ($\alpha \approx 0$). Le schéma aux différences « lisse » les données initiales.

5. Lissage de la solution d'un problème aux différences en tant qu'effet de la viscosité d'approximation. Le spectre du schéma aux différences

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad m = 0, \pm 1, \ldots \\ &\qquad\qquad\qquad\qquad\qquad\qquad p = 0, 1, \ldots, [T/\tau] - 1, \\ &\qquad\qquad\qquad\qquad\qquad\qquad m = 0, \pm 1 \\ &u_m^0 = \varphi(x_m), \end{aligned}\right\} \tag{46}$$

approchant le problème de Cauchy

$$\left.\begin{aligned} &\frac{\partial u}{\partial t}-\frac{\partial u}{\partial x}=0, \quad -\infty<x<\infty, \quad 0<t<T, \\ &u(x,0)=\varphi(x), \quad -\infty<x<\infty, \end{aligned}\right\} \tag{47}$$

est, nous l'avons vu, la circonférence $\lambda = 1-r+re^{i\alpha}$, $0 \leqslant \alpha < 2\pi$. Pour $r < 1$, il correspond à chaque point $\alpha \neq 0$ un point $\lambda(\alpha)$, $|\lambda(\alpha)| < 1$, du spectre. Cela veut dire que chaque harmonique $u_m^0 = e^{i\alpha mh}$ défini comme données initiales s'amortit en se multipliant par $\lambda(\alpha)$ lors de chaque passage d'un niveau au suivant. La solution est lissée avec le temps parce que pour αh faibles (harmoniques basses fréquences) l'amortissement est moins prononcé. Notons que la solution $u(x, t) = \varphi(x+t)$ du problème différentiel (47) n'est pas lissée avec le temps : elle s'obtient à partir des données initiales par un déplacement à gauche. Ceci étant, la solution du problème (47) correspondant à la condition initiale $u(x, 0) = e^{i\alpha x}$ est $u(x, t) = e^{i\alpha t}e^{i\alpha x}$ et le facteur $e^{i\alpha t}$ est de module un. L'effet numérique du lissage de la solution dans le cas du schéma aux différences (46) peut être assimilé à la manifestation de la *viscosité d'approximation* propre à ce schéma. Expliquons ce que nous entendons par ce terme. Si l'on estime que l'équation $\dfrac{\partial u}{\partial t}-\dfrac{\partial u}{\partial x}=0$ est le plus simple modèle de l'équation du mouvement d'un gaz non visqueux, alors il est naturel de prendre

$$\frac{\partial u}{\partial t}-\frac{\partial u}{\partial x}=\mu\frac{\partial^2 u}{\partial x^2} \tag{48}$$

pour modèle d'équations régissant le mouvement d'un gaz visqueux avec viscosité $\mu > 0$ lissant la solution. Dans la condition initiale $u(x, 0) = e^{i\alpha x}$, la solution de l'équation (48) est de la forme

$$u(x, t) = e^{-\mu\alpha^2 t+i\alpha t}e^{i\alpha x} \equiv \tilde{\lambda}(\alpha, t)e^{i\alpha x}.$$

Pour $\mu = O(\tau)$ et $t = \tau$, le facteur $\tilde{\lambda}(\alpha, t)$ atténuant l'harmonique $e^{i\alpha x}$ peut s'écrire

$$\tilde{\lambda}(\alpha, \tau) = 1-\mu\alpha^2\tau+i\alpha\tau-\frac{\alpha^2\tau^2}{2}+o(\tau^2). \tag{49}$$

Supposons que la solution $u^{(h)}$ du problème aux différences peut être prolongée en dehors du réseau de façon que la fonction régulière $u^{(h)}(x, t)$ résultante soit uniformément bornée en h ainsi que ses dérivées jusqu'à l'ordre quatre inclus.

Moyennant la formule de Taylor on écrit alors aux points (x, t) du réseau

$$\begin{aligned} 0 &= \frac{u^{(h)}(x, t+\tau)-u^{(h)}(x, t)}{\tau}-\frac{u^{(h)}(x+h, t)-u^{(h)}(x, t)}{h} = \\ &= \frac{\partial u^{(h)}(x, t)}{\partial t}-\frac{\partial u^{(h)}(x, t)}{\partial x}+\frac{\tau}{2}\frac{\partial^2 u^{(h)}(x, t)}{\partial t^2}-\frac{h}{2}\frac{\partial^2 u^{(h)}(x, t)}{\partial x^2}+ \\ &\qquad +h^2\varepsilon_1^{(h)}(x, t). \end{aligned} \tag{50}$$

Ici et dans la suite $\varepsilon_1^{(h)}$, $\varepsilon_2^{(h)}$, $\varepsilon_3^{(h)}$ sont des fonctions bornées uniformément en h, ainsi que chacune de leurs dérivées.

Il résulte en particulier de l'égalité (50)

$$\frac{\partial u^{(h)}}{\partial t} = \frac{\partial u^{(h)}}{\partial x}+h\varepsilon_2^{(h)}(x, t).$$

La dérivation par rapport à t donne

$$\frac{\partial^2 u^{(h)}}{\partial t^2} = \frac{\partial}{\partial x}\left(\frac{\partial u^{(h)}}{\partial t}\right)+h\frac{\partial\varepsilon_2^{(h)}}{\partial t} = \frac{\partial^2 u^{(h)}}{\partial x^2}+h\frac{\partial\varepsilon_2^{(h)}}{\partial x}+h\frac{\partial\varepsilon_2^{(h)}}{\partial t} = \frac{\partial^2 u^{(h)}}{\partial x^2}+h\varepsilon_3^{(h)}.$$

Portons l'expression de $\partial^2 u^{(h)}/\partial t^2$ dans (50) et rejetons les termes infiniment petits d'ordre

deux, il vient une équation différentielle du type (48) :

$$\frac{\partial U^{(h)}}{\partial t}-\frac{\partial U^{(h)}}{\partial x}=\frac{h-\tau}{2}\,\frac{\partial^2 U^{(h)}}{\partial x^2}, \tag{51}$$

que nous considérerons non pas sur le réseau, mais partout pour $t > 0$.

Ainsi, l'équation aux différences (46) « coïncide pour l'essentiel » avec l'*approximation différentielle* (51) qui est une équation de la forme (48) avec une petite viscosité $\mu = (h-\tau)/2$. Cette viscosité est dite *d'approximation* parce qu'apparue par approximation du problème (47) par le problème aux différences (46). L'équation différentielle (51) lisse les données initiales à peu près comme le schéma (46). En effet, si $U(x, 0) = e^{i\alpha x}$, à l'instant $t = \tau$ cet harmonique se trouve multiplié, conformément à la formule (49), par

$$\lambda(\alpha, \tau) = 1-\frac{h-\tau}{2}\,\alpha^2\tau+i\alpha\tau-\frac{\alpha^2\tau^2}{2}+o(\tau^2) = 1+i\alpha\tau-\frac{h}{2}\alpha^2\tau+o(\tau^2). \tag{52}$$

Lorsque $u_m^0 = e^{i\alpha x}|_{x=mh} = e^{i\alpha mh}$, on obtient à l'époque $t = \tau$ d'après (46) le même harmonique multiplié par le facteur

$$\lambda(\alpha) = 1-r+re^{i\alpha h} = 1-r+r\left(1+i\alpha h-\frac{\alpha^2h^2}{2}\right)+o(h^2) = 1+i\alpha\tau-\frac{h}{2}\,\alpha^2\tau+O(\tau^2)$$

coïncidant avec (52) à des infiniment petits d'ordre deux en τ (ou en h) près.

Exercices

1. Pour quelles valeurs du paramètre $\sigma > 0$, le schéma aux différences approchant le problème de Cauchy pour l'équation de la chaleur

$$\left.\begin{aligned}&\frac{u_m^{p+1}-u_m^p}{\tau} = \sigma\,\frac{u_{m+1}^{p+1}-2u_m^{p+1}+u_{m-1}^{p+1}}{h^2}+(1-\sigma)\,\frac{u_{m+1}^p-2u_m^p+u_{m-1}^p}{h^2},\\ &\qquad u_m^0 \quad \text{donné}, \quad m = 0, \pm 1, \ldots\end{aligned}\right\}$$

vérifie le critère spectral de stabilité de von Neumann quel que soit $r = \tau/h^2$?

2. Le critère spectral de stabilité est-il satisfait par le schéma aux différences

$$\left.\begin{aligned}&\frac{u_m^{p+1}-u_m^{p-1}}{2\tau}-\frac{u_{m+1}^p-2u_m^p+u_{m-1}^p}{h^2} = \varphi(x_m, t_p), \quad p \geqslant 1,\\ &\left.\begin{aligned}u_m^0 &= \psi_m,\\ u_m^1 &= \psi_m^{(1)},\end{aligned}\right\} m = 0, \pm 1, \ldots,\end{aligned}\right\}$$

où

$$\psi_m^{(1)} = u(x, 0)+\tau\,\frac{\partial u(x, 0)}{\partial t} = u(x_m, 0)+\tau\,\frac{\partial^2 u(x, 0)}{\partial x^2} = \psi(x_m)+\tau\psi''(x_m)?$$

Ce schéma approche à l'ordre $O(\tau^2+h^2)$ le problème de Cauchy (19) relatif à l'équation de la chaleur.

3. Montrer que le schéma aux différences

$$\left.\begin{aligned}&\frac{u_m^{p+1}-u_m^p}{\tau}+A\,\frac{u_{m+1}^{p+1}-u_{m-1}^{p+1}}{2h} = 0, \qquad m = 0, \pm 1, \ldots,\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad p = 0, 1, \ldots,\\ &u_m^0 = \psi(x_m), \qquad\qquad\qquad\qquad m = 0, \pm 1, \ldots,\end{aligned}\right\}$$

approchant à l'ordre $O(\tau+h^2)$ le problème de Cauchy

$$\left.\begin{aligned}&\frac{\partial u}{\partial t}+A\,\frac{\partial u}{\partial x} = 0, \qquad -\infty < x < \infty, \quad t > 0,\\ &u(x, 0) = \psi(x), \qquad -\infty < x < \infty,\end{aligned}\right\}$$

vérifie le critère spectral de stabilité pour tout $r = \tau/h$ et A une constante quelconque.

4. Etudier le schéma prédicteur-correcteur aux différences pour le problème de Cauchy $u_t + Au_x = 0$, $u(x, 0) = \varphi(x)$:

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} + A\,\frac{u_{m+1/2}^{p+1/2} - u_{m-1/2}^{p+1/2}}{h} = 0, \qquad m = 0, \pm 1, \ldots, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad p = 0, 1, \ldots, \\ &\qquad\qquad\qquad\qquad u_m^0 = \varphi_m, \qquad m = 0, \pm 1, \ldots, \end{aligned}\right\}$$

$A = \text{const}$, où la fonction auxiliaire $u^{p+1/2} = \left\{u_{m+1/2}^{p+1/2}\right\}$ se définit moyennant $u^p = \left\{u_m^p\right\}$ en deux étapes : on calcule d'abord $v^p = \{v_m^p\}$ en tant que solution du problème aux différences

$$\frac{v_m^{p+1/2} - u_m^p}{\tau/2} + A\,\frac{v_{m+1}^{p+1/2} - v_{m-1}^{p+1/2}}{2h} = 0, \qquad m = 0, \pm 1, \ldots,$$

puis $u^{p+1/2} = \left\{u_{m+1/2}^{p+1/2}\right\}$ par les formules

$$u_{m+1/2}^{p+1/2} = (1-\alpha)\,\frac{v_{m+1}^{p+1/2} + v_m^{p+1/2}}{2} + \alpha\,\frac{v_{m+2}^{p+1/2} + v_{m-1}^{p+1/2}}{2}.$$

Montrer que si le paramètre d'interpolation α se trouve sur le segment $0 \leqslant \alpha \leqslant 0{,}25$, alors, quel que soit $r = \tau/h = \text{const}$, le critère spectral de stabilité est valable. Lorsque $\alpha = 0$, le spectre tout entier se situe sur la circonférence unité, et pour $0 < \alpha \leqslant 0{,}25$, il est intérieur au cercle unité qu'il ne touche que pour $\lambda = 1$. A la valeur propre $\lambda = 1$ correspondent les fonctions propres $u_m = (\pm 1)^m$.

§26. Principe des coefficients fixés

Le procédé que nous allons exposer étend considérablement la classe de problèmes aux différences instationnaires dont l'étude procède du critère spectral de stabilité. Ce critère nécessaire de stabilité énoncé au §25 pour le problème de Cauchy aux différences à coefficients constants'applique également au problème de Cauchy aux différences à coefficients « continus » mais non constants, ainsi qu'aux problèmes dans des domaines bornés lorsque les conditions aux limites sont données non seulement pour $t = 0$ mais aussi sur les bords. Ce procédé peut également être utile dans l'analyse des problèmes non linéaires.

1. Fixation des coefficients aux points intérieurs. Exposons le principe des coefficients fixés sur l'exemple du problème aux limites aux différences suivant :

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - a(x_m, t_p)\,\frac{u_{m-1}^p - 2u_m^p + u_{m+1}^p}{h^2} = 0, \\ &\qquad p = 0, 1, \ldots, [T/\tau] - 1, \\ &u_m^0 = \varphi_m, \quad m = 0, 1, \ldots, M; \quad Mh = 1, \\ &\qquad l_1 u_0^{p+1} = 0, \quad l_2 u_M^{p+1} = 0. \end{aligned}\right\} \tag{1}$$

Ici $l_1 u_0^{p+1} = 0$ et $l_2 u_M^{p+1} = 0$ sont certaines conditions imposées respectivement aux points extrêmes gauche et droit du segment maillé $0 \leqslant m \leqslant M$; $a(x, t) > 0$.

Prenons un point intérieur $(\tilde{x}, \tilde{t})$ du domaine $0 < x < 1$, $0 < t < T$, où le problème (1) est proposé, et « fixons »-y les coefficients de (1).

L'équation aux différences à coefficients constants ainsi formée

$$\left.\begin{aligned} &\frac{u_m^{p+1}-u_m^p}{\tau}-a(\tilde{x}, \tilde{t})\frac{u_{m-1}^p-2u_m^p+u_{m+1}^p}{h^2} = 0, \\ &p = 0, 1, \ldots, [T/\tau]-1; \quad m = 0, \pm 1, \ldots, \end{aligned}\right\} \tag{2}$$

sera considérée désormais non pas pour $0 < m < M$ mais pour tous les m entiers. Enonçons-le

PRINCIPE DES COEFFICIENTS FIXÉS. *Pour que le problème* (1) *soit stable, il faut que le problème de Cauchy pour l'équation aux différences à coefficients constants* (2) *vérifie le critère spectral nécessaire de stabilité de von Neumann.*

Etayons ce principe par des considérations heuristiques suivantes.

Le réseau se resserrant, le coefficient $a(x, t)$ varie, vu la continuité de la fonction $a(x, t)$, de plus en plus faiblement dans le voisinage du point $(\tilde{x}, \tilde{t})$ durant tout nombre fixe des pas en espace h et des pas en temps τ et diffère de moins en moins de la valeur $a(\tilde{x}, \tilde{t})$. Notons en outre que la distance entre le point $(\tilde{x}, \tilde{t})$ et les frontières $x = 0$ et $x = 1$ du segment qui est mesurée en pas du réseau, tend vers l'infini. Si le réseau est fin, les perturbations imposées à la solution du problème (1) au temps $t = \tilde{t}$ se propagent donc dans le voisinage du point $x = \tilde{x}$ (en temps faible) à peu près comme dans le cas du problème (2).

Il s'agit certes des raisonnements généraux qui ne dépendent pas du nombre de variables d'espace, de celui de fonctions cherchées et de la forme de l'équation ou du système aux différences.

Au §25 nous avons examiné le problème de Cauchy pour une équation de la forme (2) et constaté que pour avoir la condition de von Neumann le rapport $r = \tau/h^2$ des pas du réseau doit vérifier l'inégalité

$$r \leqslant \frac{1}{2a(\tilde{x}, \tilde{t})}.$$

Le problème (1) étant stable si, en vertu du principe des coefficients fixés, la condition ci-dessus est remplie pour n'importe quel $(\tilde{x}, \tilde{t})$, le rapport $r = \tau/h$ doit vérifier l'inégalité

$$r \leqslant \frac{1}{2 \max\limits_{\tilde{x}, \tilde{t}} a(\tilde{x}, \tilde{t})}. \tag{3}$$

Grâce au principe des coefficients fixés, on reste au niveau heuristique en étudiant la stabilité des problèmes non linéaires. Voyons par exemple le problème non linéaire suivant :

$$\begin{aligned} &u_t-(1+u^2)u_{xx} = 0, && 0 < x < 1, \\ &u(x, 0) = \psi_0(x), && 0 < x < 1, \\ &u(0, t) = \psi_1(t), && u(1, t) = \psi_2(t), \quad 0 < t < T. \end{aligned}$$

Utilisons le schéma aux différences

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1}-u_m^p}{\tau_p} - [1+(u_m^p)^2]\dfrac{u_{m+1}^p-2u_m^p+u_{m-1}^p}{h^2} = 0, \\ 0 < m < M, \qquad p = 0, 1, \ldots, [T/\tau]-1, \\ u^0 = \varphi_0(mh), \qquad 0 \leqslant m \leqslant M, \\ u_0^p = \varphi_1(t_p+\tau_p), \quad p = 1, 2, \ldots, [T/\tau], \\ u_M^p = \varphi_2(t_p+\tau_p) \end{cases}$$

admettant une variation du pas τ_p d'un niveau à l'autre. Ce schéma permet de calculer successivement, niveau par niveau, u_m^1, $m = 0, \ldots, M$, u_m^2, $m = 0, 1, \ldots, M$, et ainsi de suite.

Supposons que nous avons atteint le niveau $t = t_p$, calculé u_m^p, $m = 0, 1, \ldots, M$, et que nous voulons continuer.

Comment choisir le pas suivant $\tau = \tau_p$? On peut admettre qu'il faut trouver la solution de l'équation linéaire aux différences

$$\frac{u_m^{p+1}-u_m^p}{\tau_p} - a(x_m, t_p)\frac{u_{m+1}^p-2u_m^p+u_{m-1}^p}{h^2} = 0$$

à coefficient variable donné $a(x_m, t_p) \equiv 1+(u_m^p)^2$. En effet, il est naturel de penser que les valeurs u_m^p sont voisines des valeurs $u(x_m, t_p)$ de la solution régulière $u(x, t)$ du problème différentiel. Le coefficient est alors proche de la fonction continue $a(x, t) \equiv 1+u^2(x, t)$ qui varie peu durant plusieurs pas de temps.

En appliquant le critère de von Neumann à l'équation à coefficient variable $a(x_m, t_p)$ on impose la contrainte (3) au rapport des pas du réseau qui est une condition nécessaire de stabilité :

$$\frac{\tau_p}{h^2} = r_p \leqslant \frac{1}{2\max\limits_x |a(x, t_p)|} = \frac{1}{2\max\limits_m |1+(u_m^p)^2|}.$$

On recommande donc de choisir le pas suivant τ_p à partir de la condition

$$\tau_p \leqslant \frac{1}{2\max\limits_m |1+(u_m^p)^2|} h^2.$$

L'expérience numérique sur machine confirme la justesse de ces considérations heuristiques.

Si la condition nécessaire de stabilité obtenue en considérant le problème de Cauchy à coefficients fixés en un point arbitraire du domaine est altérée, on n'aura pas de stabilité quelle que soit la façon dont on se donne les conditions aux limites. Soulignons que le principe des coefficients fixés ne tient pas compte de l'influence des conditions aux limites. La condition nécessaire de stabilité découlant de ce principe étant remplie, on a ou non la stabilité selon les conditions aux limites données. Exposons maintenant le critère de Babenko et Guelfand qui prend en considération l'influence des frontières pour un problème donné sur un segment.

2. Critère de Babenko et Guelfand. Lors de notre examen du problème (1) nous avons supposé que dans le cas d'un réseau fin les perturbations induites sur la solution du problème (1) dans le voisinage d'un point intérieur $(\bar{x}, \bar{t})$ se propagent approximativement comme les mêmes perturbations communiquées à la solution du problème de Cauchy (2) à coefficients fixés en $(\bar{x}, \bar{t})$. En justifiant ce principe nous avons admis que les distances du point intérieur $(\bar{x}, \bar{t})$ aux frontières, distances mesurées en pas du réseau, augmentent indéfiniment avec le resserrement du réseau. Mais si $(\bar{x}, \bar{t})$ appartient à la frontière $\bar{x} = 0$ ou $\bar{x} = 1$, ce raisonnement heuristique ne porte plus. Soit, par exemple, $\bar{x} = 0$. La distance entre $\bar{x}$ et tout point fixe $x > 0$ (en particulier l'extrémité droite $x = 1$ du segment) mesurée en pas du réseau augmente toujours indéfiniment pour $h \to 0$, mais le nombre de pas jusqu'à l'extrémité gauche $x = 0$ ne varie pas et reste égal à 0.

Aussi la perturbation de la solution du problème (1) au voisinage de $x = 0$ se propage-t-elle durant une petite période comme celle de la solution du problème

$$\left.\begin{aligned} \frac{u_m^{p+1} - u_m^p}{\tau} - a(0, \bar{t}) \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} &= 0, \quad m = 1, 2, \ldots, \\ l_1 u_0^{p+1} &= 0. \end{aligned}\right\} \tag{4}$$

Ce problème est obtenu à partir du problème de départ (1) en fixant le coefficient $a(x, \bar{t})$ à l'extrémité gauche $\bar{x} = 0$ du segment et en éloignant simultanément la frontière droite à $+\infty$. Le problème (4) n'est naturellement considéré que pour des fonctions $u^p = \{u_0^p, u_1^p, u_2^p, \ldots\}$ telles que

$$u_m^p \to 0 \quad \text{pour} \quad m \to +\infty.$$

C'est le seul cas où la perturbation est concentrée au voisinage de la frontière $x = 0$, et ce n'est qu'au sens de telles perturbations que le problème (1) et le problème (4) s'apparentent à proximité du point extrême gauche $x = 0$.

De même, la propagation des perturbations agissant sur la solution du problème (1) au voisinage de la frontière droite $x = 1$ doit ressembler à celle des perturbations analogues pour le problème

$$\left.\begin{aligned} \frac{u_m^{p+1} - u_m^p}{\tau} - a(1, \bar{t}) \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} &= 0, \\ m = \ldots, -2, -1, 0, 1, \ldots, M-1, \\ l_2 u_M^{p+1} &= 0 \end{aligned}\right\} \tag{5}$$

possédant seulement la frontière droite. Le problème (5) s'obtient de (1) en fixant le coefficient $a(x, \bar{t})$ à l'extrémité droite $\bar{x} = 1$ et en éloignant la frontière gauche à $-\infty$. Le problème (5) est à considérer pour des fonctions discrètes $u^p = \{\ldots, u_{-2}^p, u_{-1}^p, u_0^p, u_1^p \ldots, u_M^p\}$ telles qu'on ait $u_m^p \to 0$ pour $m \to -\infty$.

Les problèmes (2), (4) et (5) sont plus simples que le problème de départ (1) en ce sens que pour $r = \tau/h^2$ fixe ils ne dépendent pas de h et sont à coefficients constants.

En ce qui concerne le problème (1), l'étude de la stabilité compte tenu de l'influence des frontières se fait donc selon le schéma suivant. On forme trois problèmes auxiliaires (2), (4) et (5) indépendants de h pour chacun desquels on trouve tous les nombres λ (valeurs propres de l'opérateur $u^p \to u^{p+1}$) tels qu'il existe des solutions de la forme

$$u_m^p = \lambda^p u_m^0.$$

Ceci étant, pour le problème (2) $u^0 = \{u_m^0\}$, $m = 0, \pm 1, \ldots$, doit être borné; pour (4) $u^0 = \{u_0^0, u_1^0, \ldots, u_m^0, \ldots\}$, $u_m^0 \to 0$ avec $m \to +\infty$; pour (5)

$$u^0 = \{\ldots, u_{-2}^0, u_{-1}^0, u_0^0, u_1^0, \ldots, u_M^0\},$$
$$u_m^0 \to 0 \quad \text{pour} \quad m \to -\infty.$$

Pour que le problème (1) *soit stable, il faut que l'ensemble de valeurs propres de chacun des trois problèmes* (2), (4) *et* (5) *se trouve dans le cercle unité* $|\lambda| \leqslant 1$. (*Le problème* (2) *est considéré pour chaque* $\tilde{x}$ *fixe*, $0 < \tilde{x} < 1$.)

Poursuivons l'étude de (1) et supposons partout dans la suite que $a(x, t) \equiv 1$. Calculons les spectres des (2), (4) et (5) pour les différentes conditions aux limites $l_1 u_0^{p+1} = 0$ et $l_2 u_M^{p+1} = 0$.

En portant la solution $u_m^p = \lambda^p u_m$ dans l'équation aux différences (2) on obtient

$$(\lambda - 1)u_m - r(u_{m+1} - 2u_m + u_{m-1}) = 0, \quad r = \frac{\tau}{h^2},$$

ou

$$u_{m+1} - \left(\frac{-2r+1-\lambda}{r}\right) u_m + u_{m-1} = 0. \tag{6}$$

Nous avons déjà abordé de telles équations du second ordre dans le chapitre 1. Pour écrire la solution générale de (6) formons l'équation caractéristique

$$q^2 + \left(2 + \frac{\lambda - 1}{r}\right) q + 1 = 0. \tag{7}$$

Si q en est racine, la fonction discrète

$$u_m^p = \lambda^p q^m$$

est une des solutions de

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0.$$

Si $|q| = 1$, i.e. $q = e^{i\alpha}$, la fonction

$$u_m^p = \lambda^p e^{i\alpha m}$$

bornée pour $m \to \pm\infty$ est solution lorsque

$$\lambda = 1 - 4r \sin^2 \frac{\alpha}{2}, \quad 0 \leqslant \alpha \leqslant 2\pi$$

(cf. §25). Ces $\lambda = \lambda(\alpha)$ remplissent le segment $1 - 4r \leqslant \lambda \leqslant 1$ de l'axe réel, qui est justement le spectre du problème (2). Celui-ci ne possède pas

de valeurs propres λ en dehors de ce segment, puisque si l'équation caractéristique (7) n'a pas de racine q de module égal à l'unité, le problème (6) n'admet pas de solution bornée pour $m \to \pm\infty$.

Si λ est extérieur au segment $1-4r \leqslant \lambda \leqslant 1$, l'équation caractéristique (7) a ses deux racines de module différent de 1 et telles que leur produit vaille le terme libre de l'équation du second degré (7), i.e. l'unité. Aussi l'une des racines de (7) est-elle inférieure à 1 et l'autre supérieure à 1 en module. Supposons, pour fixer les idées, que $|q_1| < 1$ et $|q_2| > 1$. La solution générale de (6) de module décroissant pour $m \to +\infty$ est alors de la forme

$$u_m = c[q_1(\lambda)]^m$$

et celle qui tend vers zéro pour $m \to -\infty$

$$u_m = c[q_2(\lambda)]^m.$$

On définit les valeurs propres du problème (4) en portant $u_m = cq_1^m(\lambda)$ dans la condition à l'extrémité gauche $l_1u = 0$ et en trouvant tous les λ pour lesquels elle est remplie. Si, par exemple,

$$l_1u_0 \equiv u_0 = 0,$$

la condition $cq_1^0 = 0$ n'est vérifiée pour aucun $c \neq 0$, de sorte qu'il n'y a pas de valeurs propres.

Si $l_1u_0 = u_1-u_0 = 0$, étant donné que $q_1 \neq 1$, la condition cq_1^1- $-cq_1^0 = c(q_1-1) = 0$ conduit à $c = 0$, d'où l'absence de valeurs propres.

Dans le cas $l_1u_0 = 2u_1-u_0 = 0$, la condition $c(2q_1-1) = 0$ est satisfaite pour $c \neq 0$ si $q_1 = 1/2$.

On trouve à partir de l'équation (7) que si $q_1 = 1/2$, le nombre λ est

$$\lambda = 1+r\left(q_1-2+\frac{1}{q_1}\right) = 1+r\frac{1-4+4}{2} = 1+\frac{r}{2}.$$

C'est l'unique valeur propre du problème (4). Elle est extérieure au cercle unité parce que $\lambda = 1+r/2 > 1$. Les valeurs propres du problème (5) se calculent d'une façon analogue. Elles s'obtiennent de l'équation

$$l_2u_M = 0,$$

pour

$$u_m = q_2^m, \quad q_2 = q_2(\lambda), \quad m = M,\ M-1,\ M-2, \ldots$$

Etudions le critère de Babenko-Guelfand dans le cas du schéma aux différences

$$\left.\begin{aligned} \frac{u_m^{p+1}-u_m^p}{\tau} - \frac{u_{m+1}^p-u_m^p}{h} = 0, \qquad & p = 0, 1, \ldots, [T/\tau]-1, \\ & m = 0,\ 1,\ \ldots,\ M-1, \\ & Mh = 1, \\ u_m^0 = \psi(x_m), \quad & m = 0, 1, \ldots, M, \\ u_M^{p+1} = 0, & \end{aligned}\right\} \tag{8}$$

approchant le problème

$$\left.\begin{aligned} u_t - u_x &= 0, \qquad 0 < x < 1, \quad 0 < t < T, \\ u(x, 0) &= \psi(x), \\ u(1, t) &= 0. \end{aligned}\right\}$$

Confrontons (8) avec un problème sans frontières

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad m = 0, \pm 1, \ldots, \tag{9}$$

avec un autre à la seule frontière gauche

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad m = 0, 1, \ldots, \tag{10}$$

et un troisième à la seule frontière droite

$$\left.\begin{aligned} \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} &= 0, \quad m = M-1, M-2, \ldots, \\ u_M^{p+1} &= 0, \quad p = 0, 1, \ldots, [T/\tau]-1. \end{aligned}\right\} \tag{11}$$

Dans le cas du problème (10) il n'y a pas de condition aux limites du moment qu'il en était ainsi du problème de départ (8).

Il faut trouver l'ensemble de valeurs propres des trois opérateurs $u^p \to u^{p+1}$ correspondant à chacun des problèmes auxiliaires (9), (10), (11) et établir les conditions de leur appartenance au cercle $|\lambda| \leqslant 1$.

En substituant une solution de la forme

$$u_m^p = \lambda^p u_m$$

dans l'équation aux différences

$$u_m^{p+1} = (1-r)u_m^p + ru_{m+1}^p, \quad r = \frac{\tau}{h},$$

on aboutit à l'équation aux différences ordinaire du premier ordre pour la fonction propre :

$$(\lambda - 1 + r)u_m - ru_{m+1} = 0. \tag{12}$$

L'équation caractéristique associée

$$(\lambda - 1 + r) - rq = 0 \tag{13}$$

établit la relation entre λ et q. L'équation (12) a comme solution générale

$$u_m = cq^m = c\left(\frac{\lambda - 1 + r}{r}\right)^m, \quad m = 0, \pm 1, \ldots$$

Pour $|q| = 1$, $q = e^{i\alpha}$, $0 \leqslant \alpha \leqslant 2\pi$,

$$\lambda = (1-r) + re^{i\alpha}.$$

Le point $\lambda = \lambda(\alpha)$ parcourt une circonférence de centre le point $1-r$ et de rayon r. Ce sont précisément les valeurs propres du problème (9)

(fig. 26, *a*). Quel que soit q, $|q| < 1$, il existe une solution non triviale

$$u_m^p = \lambda^p u_m = c_0 \lambda^p q^m$$

du problème (10), qui décroît avec $m \to +\infty$.

Les valeurs propres $\lambda = 1-r+rq$ remplissent évidemment tout l'intérieur du cercle de circonférence $\lambda = (1-r)+re^{i\alpha}$ (fig. 26, *b*).

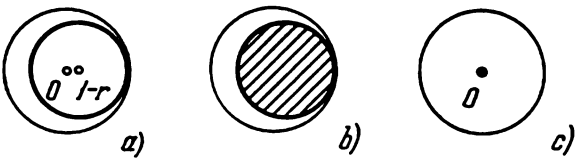


Fig. 26

Enfin, les solutions $u_m^p = \lambda^p u_m$ du problème (11) qui décroissent avec $m \to -\infty$ doivent être de la forme

$$u_m^p = c\lambda^p q^m, \quad |q| > 1,$$

λ et q étant reliés par l'égalité (13).

La condition aux limites $u_M^p = 0$ entraîne l'existence d'une solution non triviale ($c \neq 0$) seulement pour $\lambda = \lambda(q) = 0$, i.e. pour $q = (r-1)/r$. Cette quantité q est de module supérieur à l'unité si l'on a $(r-1)/r > 1$ ou $(r-1)/r < -1$. La première inégalité n'admet pas de solution, et la solution de la seconde est $r < 1/2$.

Ainsi, pour $r < 1/2$, le problème (10) a une valeur propre $\lambda = 0$ (fig. 26, *c*). Les fig. 27, *a*, *b*, *c* représentent la réunion des valeurs propres des trois problèmes pour les cas respectifs $r < 1/2$, $1/2 < r < 1$ et $r > 1$.

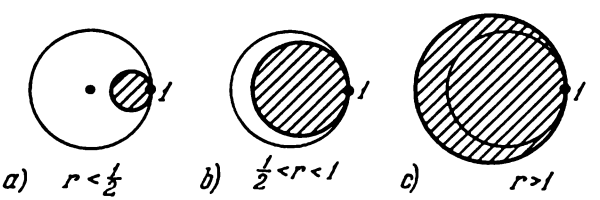


Fig. 27

Il est clair que la réunion des valeurs propres des trois problèmes est dans le cercle $|\lambda| < 1+c\tau$, avec c indépendant de h, si et seulement si $r \leqslant 1$.

Le critère exposé joue également dans le cas des problèmes aux limites sur un segment pour des systèmes d'équations aux différences. De nombreux schémas vérifiant le critère de von Neumann, schémas qui paraissent naturels de prime abord, s'avèrent alors instables à cause d'une approximation malheureuse des conditions aux limites. On voit donc à quel point il est important de savoir choisir des schémas dépourvus de ce défaut.

Dans le chap. 13 nous discuterons le critère spectral de Babenko-Guelfand des positions plus générales. Nous démontrerons en particulier de façon rigoureuse qu'il s'agit là d'une condition nécessaire de stabilité et que si elle est satisfaite, aucune violation « grossière » de la stabilité n'est possible.

Exercices

1. Dans quelles conditions a-t-on le critère spectral de stabilité pour le schéma aux différences

$$\frac{u_m^{p+1}-u_m^p}{\tau}-\frac{u_{m+1}^p-u_{m-1}^p}{2h}-\frac{\tau}{2}\,\frac{u_{m+1}^p-2u_m^p+u_{m-1}^p}{h^2}=0,$$

$$m=1,2,\ldots,\quad M-1,$$

$$u_m^0=\varphi(x_m),\qquad m=0,1,\ldots,M,$$

$$\left.\begin{aligned} u_M^{p+1}&=0,\\ \frac{u_0^{p+1}-u_0^p}{\tau}-\frac{u_1^p-u_0^p}{h}&=0,\end{aligned}\right\}\quad p=0,1,\ldots,[T/\tau]-1,$$

approchant le problème différentiel

$$\left.\begin{aligned} &u_t-u_x=0,\qquad 0<x<1,\quad 0<t<T,\\ &u(x,0)=\varphi(x),\\ &u(0,t)=u(1,t)=0\end{aligned}\right\}$$

pour la solution régulière $u(x, t)$ à l'ordre deux en h ?

Réponse : $\tau/h \leqslant 1$.

2. Pour construire un schéma aux différences approchant le problème aux limites suivant relatif au système hyperbolique d'équations différentielles

$$\left.\begin{aligned}\frac{\partial v}{\partial t}&=\frac{\partial w}{\partial x},\\ \frac{\partial w}{\partial t}&=\frac{\partial v}{\partial x},\end{aligned}\right.\qquad 0\leqslant x\leqslant 1,\quad 0\leqslant t\leqslant T,$$

$$v(x,0)=\varphi_1(x),\qquad w(x,0)=\varphi_2(x),$$

$$v(0,t)=w(1,t)=0,$$

posons $\mathbf{u}(x, t)=\begin{pmatrix}v(x,t)\\ w(x,t)\end{pmatrix}$ et mettons le problème sous forme matricielle

$$\frac{\partial}{\partial t}\mathbf{u}-A\frac{\partial}{\partial x}\mathbf{u}=0,$$

$$\mathbf{u}(x,0)=\boldsymbol{\phi}(x),$$

$$v(0,t)=w(1,t)=0,$$

avec $A=\begin{pmatrix}0&1\\1&0\end{pmatrix}$. Choisissons le réseau $(x_m, t_n)=(mh, n\tau)$, $h=1/M$, M entier naturel.

Posons

$$\frac{u_m^{p+1}-u_m^p}{\tau}-A\,\frac{u_{m+1}^p-u_{m-1}^p}{2h}-\frac{\tau}{2}A^2\,\frac{u_{m+1}^p-2u_m^p+u_{m-1}^p}{h^2}=0,$$

$$m=1,2,\ldots,M-1,$$

$$u_m^0=\phi(x_m),$$

$$v_0^{p+1}=w_M^{p+1}=0.$$

Notre construction s'achève avec la donnée des conditions supplémentaires aux bords de gauche et de droite. En notant que, quels que soient α et β, le système d'équations différentielles entraîne les égalités

$$\left.\frac{\partial(v+\alpha w)}{\partial t}-\frac{\partial(w+\alpha v)}{\partial x}\right|_{x=0}=0,$$

$$\left.\frac{\partial(v+\beta w)}{\partial t}-\frac{\partial(w+\beta v)}{\partial x}\right|_{x=1}=0,$$

donnons-nous des conditions supplémentaires en posant

$$\frac{(v_0^{p+1}+\alpha w_0^{p+1})-(v_0^p+\alpha w_0^p)}{\tau}-\frac{(w_1^p+\alpha v_1^p)-(w_0^p+\alpha v_0^p)}{h}=0,$$

$$\frac{(v_M^{p+1}+\beta w_M^{p+1})-(v_M^p+\beta w_M^p)}{\tau}-\frac{(w_M^p+\beta v_M^p)-(w_{M-1}^p+\beta v_{M-1}^p)}{h}=0.$$

Sous l'hypothèse $r=\tau/h\leqslant 1$ montrer que

a) si $\alpha=1$, $\beta=-1$, le critère spectral de stabilité est satisfait ;

b) si $\alpha=-1$, le critère n'est pas satisfait quel que soit β.

c) Quelles doivent être les conditions imposées à α et β pour qu'ait lieu le critère spectral de stabilité tenant compte de l'influence des conditions aux limites ?

§27. Solutions de certains problèmes modèles sous forme des séries finies de Fourier

Citons plusieurs problèmes modèles dont les solutions sont représentables sous forme des séries finies de Fourier. De telles représentations sont extrêmement utiles. En effet, elles permettent de pénétrer les propriétés des problèmes modèles considérés et, partant, celles de la classe de problèmes réels correspondants.

Voyons pour commencer ce qu'est la série de Fourier pour une fonction définie sur un réseau.

1. Séries de Fourier pour des fonctions définies sur un réseau. Considérons l'ensemble de toutes les fonctions réelles $v=\{v_m\}$ définies aux points $x_m=mh$, $m=0,1,\ldots,M$; $Mh=1$, et s'annulant pour $m=0$ et $m=M$. La famille de ces fonctions munie des opérations usuelles d'addition et de multiplication par des nombres réels constitue un espace vectoriel de dimension $M-1$, puisque le système de fonctions

$$\tilde{\psi}_m^{(k)}=\begin{cases}0 & \text{si} \quad m\neq k,\\ 1 & \text{si} \quad m=k,\end{cases}$$

$k=1,2,\ldots,M-1$, forme évidemment une base. En effet, toute fonction $v=(v_0,v_1,\ldots,v_M)$, $v_0=v_M=0$, se met d'une façon unique sous forme de

combinaison linéaire de fonctions $\tilde{\psi}^{(1)}, \tilde{\psi}^{(2)}, \ldots, \tilde{\psi}^{(M-1)}$:

$$v = v_1\tilde{\psi}^{(1)} + \ldots + v_{M-1}\tilde{\psi}^{(M-1)}.$$

Munissons l'espace considéré du produit scalaire en posant

$$(v, w) = h \sum_{m=0}^{M} v_m w_m. \tag{1}$$

Montrons que le système de fonctions

$$\psi^{(k)} = \left\{\sqrt{2} \sin \frac{k\pi m}{M}\right\}, \quad k = 1, 2, \ldots, M-1, \tag{2}$$

constitue une base orthonormale dans l'espace donné, i.e.

$$(\psi^{(k)}, \psi^{(r)}) = \begin{cases} 0, & k \neq r, \\ 1, & k = r, \end{cases}$$
$$k, r = 1, 2, \ldots, M-1. \tag{3}$$

La démonstration s'effectue en remarquant que

$$\sum_{m=0}^{M-1} \cos \frac{l\pi m}{M} = \frac{1}{2} \sum_{m=0}^{M-1} \left(e^{i\frac{l\pi m}{M}} + e^{-i\frac{l\pi m}{M}}\right) = \frac{1}{2} \frac{1-e^{il\pi}}{1-e^{i\frac{l\pi}{M}}} + \frac{1}{2} \frac{1-e^{-il\pi}}{1-e^{-i\frac{l\pi}{M}}} =$$
$$= \begin{cases} 0 & \text{si } l \text{ est pair et } 0 < l < 2M, \\ 1 & \text{si } l \text{ est impair.} \end{cases}$$

D'où pour $k \neq r$

$$(\psi^{(k)}, \psi^{(r)}) = 2h \sum_{m=0}^{M} \sin \frac{k\pi m}{M} \sin \frac{r\pi m}{M} = 2h \sum_{m=0}^{M-1} \sin \frac{k\pi m}{M} \sin \frac{r\pi m}{M} =$$
$$= h \sum_{m=0}^{M-1} \cos \frac{(k-r)\pi m}{M} - h \sum_{m=0}^{M-1} \cos \frac{(k+r)\pi m}{M} = 0,$$

et pour $k = r$

$$(\psi^{(k)}, \psi^{(k)}) = h \sum_{m=0}^{M-1} \cos 0 - h \sum_{m=0}^{M-1} \cos \frac{2k\pi m}{M} = hM - h \cdot 0 = 1.$$

Toute fonction discrète $v = (v_0, v_1, \ldots, v_M)$ se développe dans la base orthonormale (2) en somme

$$v = c_1\psi^{(1)} + \ldots + c_{M-1}\psi^{(M-1)}$$

ou

$$v_m = \sqrt{2} \sum_{k=1}^{M-1} c_k \sin \frac{k\pi m}{M}, \tag{4}$$

où

$$c_k = (v, \psi^{(k)}) = \sqrt{2}h \sum_{m=0}^{M} v_m \sin \frac{k\pi m}{M}.$$

La base (2) étant orthonormale, on a évidemment

$$(v, v) = c_1^2 + c_2^2 + \ldots + c_{M-1}^2. \tag{5}$$

La somme (4) est la représentation de la fonction discrète $v = \{v_m\}$ sous forme de série finie de Fourier et l'égalité (5) l'analogue exact de l'égalité de Parseval en théorie classique des séries de Fourier.

On considère de même les séries finies de Fourier pour des fonctions définies sur un réseau carré. Soit le réseau

$$x_m = mh, \quad y_n = nh, \quad 0 \leqslant mh \leqslant 1, \quad 0 \leqslant nh \leqslant 1,$$

$h = 1/M$, M entier naturel. La famille de fonctions réelles $v = \{v_{mn}\}$ définies aux points du réseau et s'annulant sur sa frontière forme un espace vectoriel. Munissons-le du produit scalaire

$$(v, w) = h^2 \sum_{n=0}^{M} \sum_{m=0}^{M} v_{mn} w_{mn}.$$

Dans l'espace vectoriel considéré de dimension $(M-1)^2$ le système de fonctions

$$\psi^{(k,l)} = 2 \sin \frac{k\pi m}{M} \sin \frac{l\pi n}{M}, \qquad \begin{array}{l} k = 1, 2, \ldots, M-1, \\ l = 1, 2, \ldots, M-1 \end{array}$$

forme une base orthonormale

$$(\psi^{(k,l)}, \psi^{(r,s)}) = \begin{cases} 0 & \text{si} \quad k \neq r \quad \text{ou} \quad l \neq s, \\ 1 & \text{si} \quad k = r \quad \text{et} \quad l = s. \end{cases}$$

Cela résulte de (3) à condition de remarquer que

$$(\psi^{(k,l)}, \psi^{(r,s)}) = \left(2 \sum_{m=0}^{M} \sin \frac{k\pi m}{M} \sin \frac{r\pi m}{M}\right) \left(2 \sum_{n=0}^{M} \sin \frac{l\pi n}{M} \sin \frac{s\pi n}{M}\right) =$$
$$= (\psi^{(k)}, \psi^{(r)}) (\psi^{(l)}, \psi^{(s)}).$$

Toute fonction $v = \{v_{mn}\}$ s'annulant sur la frontière du carré est développable en série finie bidimensionnelle de Fourier

$$v_{mn} = 2 \sum_{k,l=1}^{M-1} c_{kl} \sin \frac{k\pi m}{M} \sin \frac{l\pi n}{M}, \tag{6}$$

où

$$c_{kl} = (v, \psi^{(k,l)}).$$

On a l'égalité de Parseval

$$(v, v) = \sum_{k,l=1}^{M-1} c_{kl}^2. \tag{7}$$

Dans tous les exemples de problèmes aux limites aux différences dont les solutions se mettront sous forme de séries finies de Fourier, on utilise

l'expression

$$\Lambda_{xx}v_m \equiv \frac{1}{h^2}(v_{m+1}-2v_m+v_{m-1}), \quad m = 1, 2, \ldots, M-1. \tag{8}$$

Notons que

$$\Lambda_{xx}\sin\frac{k\pi m}{M} = \frac{1}{h^2}\left[\sin\frac{k\pi(m+1)}{M}-2\sin\frac{k\pi m}{M}+\sin\frac{k\pi(m-1)}{M}\right] =$$

$$= \frac{2}{h^2}\left(\cos\frac{k\pi}{M}-1\right)\sin\frac{k\pi m}{M} = \mu_k\sin\frac{k\pi m}{M}, \quad m = 1, 2, \ldots, M-1,$$

avec

$$\mu_k = -\frac{4}{h^2}\sin^2\frac{k\pi}{2M}.$$

En d'autres termes, la base (2) est formée de fonctions propres de l'opérateur Λ_{xx} transformant $v = \{v_m\}$ de l'espace de fonctions s'annulant pour $m = 0$ et $m = M$ en des éléments du même espace par les formules

$$w_m = \frac{1}{h^2}(v_{m+1}-2v_m+v_{m-1}), \quad m = 1, 2, \ldots, M-1.$$

Il correspond à la fonction propre $\psi^{(k)} = \sqrt{2}\sin\frac{k\pi m}{M}$ la valeur propre

$$\mu_k = -\frac{4}{h^2}\sin^2\frac{k\pi}{2M}, \quad k = 1, 2, \ldots, M-1. \tag{9}$$

2. Représentation des solutions de schémas aux différences pour l'équation de la chaleur sur un segment. Notre premier exemple où nous réussirons à mettre la solution sous forme de série finie de Fourier est le schéma aux différences très simple

$$\left.\begin{aligned} \frac{u_m^{p+1}-u_m^p}{\tau} - \frac{u_{m+1}^p-2u_m^p+u_{m-1}^p}{h^2} &= 0, \quad m = 1, 2, \ldots, M-1, \\ &\qquad p = 0, 1, \ldots, [T/\tau]-1, \\ u_0^p = u_M^p &= 0, \\ u_m^0 &= \psi(mh) \end{aligned}\right\} \tag{10}$$

pour le problème de la chaleur sur un segment

$$\left.\begin{aligned} u_t-u_{xx} &= 0, \quad 0 \leqslant t \leqslant T, \quad 0 \leqslant x \leqslant 1, \\ u(0, t) &= u(1, t) = 0, \quad 0 \leqslant t \leqslant T, \\ u(x, 0) &= \psi(x), \quad 0 \leqslant x \leqslant 1. \end{aligned}\right\} \tag{11}$$

Récrivons le problème (10):

$$\left.\begin{aligned} u_m^{p+1} &= u_m^p+\tau\Lambda_{xx}u_m^p = (E+\tau\Lambda_{xx})u_m^p, \\ u_m^0 &= \psi(mh). \end{aligned}\right\} \tag{12}$$

Ici E est une application identique: $Eu_m^p = u_m^p$, et $E+\tau\Lambda_{xx}$ l'opérateur $u^p \to u^{p+1}$ ou l'opérateur de passage d'un niveau à l'autre. Les fonctions

discrètes $u^p \equiv \{u_m^p\}$ de l'argument m sont supposées appartenir, pour chaque p fixe, à l'espace considéré, i.e. $u_0^p = u_M^p = 0$.

Cherchons les solutions de l'équation (12) sous la forme

$$\lambda_k^p \psi^{(k)} \equiv \lambda_k^p \left(\sqrt{2} \sin \frac{k\pi m}{M}\right).$$

Substituons cette expression dans (12) et simplifions les deux membres par $\lambda_k^p \sqrt{2} \sin \frac{n\pi m}{M}$; nous obtenons, en vertu de (9), l'expression suivante de λ_k:

$$\lambda_k = 1+\tau\mu_k = 1-\frac{4\tau}{h^2} \sin^2 \frac{k\pi}{2M}, \quad k = 1, 2, \ldots, M-1.$$

La linéarité de (12) fait que l'expression

$$u^p = \sum_{k=1}^{M-1} c_k \lambda_k^p \psi^{(k)} \tag{13}$$

est sa solution quelles que soient les constantes c_k. Lorsque $p = 0$,

$$u_m^0 = \sum_{k=1}^{M-1} c_k \left(\sqrt{2} \sin \frac{k\pi m}{M}\right).$$

Prenons pour c_k les coefficients du développement de la fonction donnée $u_m^0 = \psi(mh)$ en série finie de Fourier, i.e. posons

$$c_k = (\psi, \psi^{(k)}) = h \sum_{m=0}^{M} \psi(mh) \left(\sqrt{2} \sin \frac{k\pi m}{M}\right).$$

La solution (13)

$$u_m^p = \sum_{k=1}^{M-1} c_k \left(1-\frac{\tau}{4h^2} \sin^2 \frac{\pi k}{2M}\right)^p \left(\sqrt{2} \sin \frac{k\pi m}{M}\right) \tag{14}$$

vérifie alors la condition initiale donnée $u_m^0 = \psi(mh)$. La formule (14) est bien la représentation cherchée de la solution sous forme de série finie de Fourier.

Les coefficients $c_k^{(p)}$ du développement de la fonction $u_m^{(p)}$ d'argument m dans la base orthonormale $\psi^{(k)} = \sqrt{2} \sin \frac{k\pi m}{M}$

$$u_m^p = \sum c_k^{(p)} \psi^{(k)}$$

s'écrivent pour p fixe

$$c_k^{(p)} = c_k \lambda_k^p.$$

Etant donné l'égalité de Parseval, on a donc

$$(u^{p+1}, u^{p+1}) = \sum_{k=1}^{M-1} |c_k^{(p+1)}|^2 = \sum_{k=1}^{M-1} |c_k \lambda_k^{p+1}|^2 \leqslant$$

$$\leqslant \max_k |\lambda_k|^2 \sum_{k=1}^{M-1} |c_k \lambda_k^p|^2 = \max_k |\lambda_k|^2 (u^p, u^p),$$

l'égalité stricte $(u^{p+1}, u^{p+1}) = \max_k |\lambda_k|^2 (u^p, u^p)$ étant réalisée si en qualité de u^0 on prend $\psi^{(k)}$ de $|\lambda_k|$ maximal.

Si $\max_k |\lambda_k|^2 \leqslant 1$, on a l'inégalité

$$(u^{p+1}, u^{p+1}) \leqslant (u^p, u^p). \tag{15}$$

Les formes quadratiques définies positives

$$(Au^p, u^p),$$

avec A la matrice de la forme quadratique, rappellent les expressions de l'énergie de la physique mathématique. Les inégalités de la forme

$$(Au^{p+1}, u^{p+1}) \leqslant (Au^p, u^p)$$

utilisées pour la résolution des problèmes aux limites aux différences s'appellent donc généralement *inégalités de l'énergie.*

Ainsi, l'estimation (15) constitue une inégalité de l'énergie parmi les plus simples. Si l'inégalité de l'énergie est satisfaite, il est naturel de choisir les normes $\|\cdot\|_{\sigma_h}$ et $\|\cdot\|_{F_h}$ en fonction de la forme (Au^p, u^p) en posant en particulier $\|u^{(h)}\|_{\sigma_h} = \max_p (Au^p, u^p)^{1/2}$. De telles normes sont dites *normes énergie.*

L'inégalité $\max_k |\lambda_k|^2 \leqslant 1$ est évidemment vraie si

$$r = \frac{\tau}{h^2} \leqslant \frac{1}{2}.$$

Lorsque

$$r = \text{const} > \frac{1}{2}$$

et pour des h suffisamment petits, il existe des λ_k tels que $|\lambda_k| > 1$. Aucun choix raisonnable (*) des normes ne garantit alors la stabilité.

Soit le schéma aux différences plus général

$$\frac{u_m^{p+1} - u_m^p}{\tau} - [(1-\sigma)\Lambda_{xx}u^p + \sigma\Lambda_{xx}u^{p+1}] = 0,$$

$$u_m^0 = \psi(mh)$$

pour le même problème différentiel de la chaleur (11). Ici σ est un paramètre.

Trouvons des solutions de la forme

$$u_m^p = \lambda_k^p \sqrt{2} \sin \frac{k\pi m}{M}, \quad k = 1, 2, \ldots, M-1,$$

avec λ_k à définir.

Portons cette expression dans l'équation aux différences ; nous obtenons

(*) Voir §13.

une relation satisfaite par λ_k :

$$\lambda_k = 1+\tau(1-\sigma)\mu_k+\tau\sigma\lambda_k\mu_k.$$

D'où

$$\lambda_k = \frac{1-\dfrac{(1-\sigma)\tau}{4h^2}\sin^2\dfrac{k\pi}{2M}}{1+\dfrac{\sigma\tau}{4h^2}\sin^2\dfrac{k\pi}{2M}}, \quad k = 1, 2, \ldots, M-1.$$

On a toujours

$$(u^{p+1}, u^{p+1}) \leqslant \max_k |\lambda_k|^2 (u^p, u^p).$$

L'inégalité de l'énergie (15) est juste dans la condition

$$\max |\lambda_k| \leqslant 1$$

ou

$$\left|1-\frac{(1-\sigma)r}{4}\sin^2\frac{k\pi}{2M}\right| \leqslant \left|1+\frac{\sigma r}{4}\sin^2\frac{k\pi}{2M}\right|, \quad r = \frac{\tau}{h^2}.$$

Il est évident que pour $1 \geqslant \sigma \geqslant 1/2$, cette inégalité, de même que (15), a lieu quel que soit r. Si $\sigma = 0$, le schéma aux différences change en schéma explicite déjà examiné, et, comme nous l'avons vu, pour avoir (15) avec n'importe quel h, il faut que $r \leqslant 1/2$.

3. Représentation des solutions de schémas aux différences pour le problème plan de la chaleur. Considérons le problème de la chaleur à deux dimensions

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}, \quad 0 \leqslant x \leqslant 1, \quad 0 \leqslant y \leqslant 1, \quad 0 \leqslant t \leqslant T. \\ &u(x, y, 0) = \psi(x, y), \\ &u(x, y, t)|_\Gamma = 0, \end{aligned}\right\} \tag{16}$$

Γ désignant ici la surface latérale du parallélépipède $0 \leqslant x, y \leqslant 1, 0 \leqslant t \leqslant T$.

Construisons le réseau $(x_m, y_n, t_p) = (mh, nh, p\tau)$ et admettons que $h = 1/M$, M entier naturel. Acceptons en qualité de D_h les points du réseau à l'intérieur et sur le contour du parallélépipède $0 \leqslant x, y \leqslant 1, 0 \leqslant t \leqslant T$.

Notons

$$\Lambda_{xx}u^p_{mn} = \frac{u^p_{m+1,n}-2u^p_{mn}+u^p_{m-1,n}}{h^2},$$

$$\Lambda_{yy}u^p_{mn} = \frac{u^p_{m,n+1}-2u^p_{mn}+u^p_{m,n-1}}{h^2}.$$

Les opérateurs Λ_{xx} et Λ_{yy} sont complètement analogues, mais le premier agit sur la variable m et a n et p pour paramètres et le second opère sur n, m et p étant des paramètres.

Le plus simple schéma aux différences pour le problème (16) est

$$\left.\begin{aligned} \frac{u_{mn}^{p+1}-u_{mn}^{p}}{\tau} &= \Lambda_{xx}u_{mn}^{p}+\Lambda_{yy}u_{mn}^{p}, && 0 \leqslant mh, nh \leqslant 1, \\ u_{mn}^{0} &= \psi(mh, nh), && 0 \leqslant p\tau < T-\tau, \\ u_{mn}^{p}|_{\Gamma} &= 0. \end{aligned}\right\} \tag{17}$$

Cherchons les solutions de l'équation aux différences dans la condition $u_{mn}^{p}|_{\Gamma} = 0$ sous la forme

$$u_{mn}^{p} = \lambda_{kl}^{p}\psi_{mn}^{(k,l)}.$$

Notons que

$$\Lambda_{xx}\psi^{(k,l)} = \Lambda_{xx}(\psi_m^{(k)}\psi_n^{(l)}) = \psi^{(l)}\Lambda_{xx}\psi^{(k)} = \mu^{(k)}\psi^{(k)}\psi^{(l)} = \mu^{(k)}\psi^{(k,l)},$$
$$\Lambda_{yy}\psi^{(k,l)} = \Lambda_{yy}(\psi_m^{(k)}\psi_n^{(l)}) = \mu^{(l)}\psi^{(k,l)}.$$

Pour λ_{kl} on a donc l'expression

$$\frac{\lambda_{kl}-1}{\tau} = [\mu^{(k)}+\mu^{(l)}]$$

ou

$$\lambda_{kl} = 1-\frac{4\tau}{h^2}\left(\sin^2\frac{k\pi}{2M}+\sin^2\frac{l\pi}{2M}\right).$$

La solution

$$u^{p} = \sum_{k,l=1}^{M-1} c_{kl}\lambda_{kl}^{p}\psi^{(k,l)} \tag{18}$$

vérifie les conditions sur la surface latérale pour tout choix des constantes c_{kl}. Pour $p = 0$, cette solution s'écrit

$$u^{0} = \sum c_{kl}\psi^{(k,l)}.$$

Pour que la condition initiale donnée

$$u_{mn}^{0} = \psi(mh, nh) = \sum c_{kl}\psi_{mn}^{(k,l)}$$

soit remplie, il faut prendre pour c_{kl} les coefficients de Fourier de la fonction $\psi(mh, nh)$, i.e.

$$c_{kl} = h^2\sum_{m,n=0}^{M}\psi(mh, nh)\left(2\sin\frac{k\pi m}{M}\sin\frac{l\pi m}{M}\right). \tag{19}$$

En vertu de (18), le coefficient de $\psi^{(k,l)}$ dans le développement de u^{p} en série de Fourier est le nombre $(c_{kl}\lambda_{kl}^{p})$. Aussi

$$(u^{p}, u^{p}) = \sum_{k,l}|c_{kl}\lambda_{kl}^{p}|^2.$$

Vu cette égalité, on écrit pour tout p fixe

$$(u^{p+1}, u^{p+1}) = \sum_{k,l=1}^{M-1}|c_{kl}\lambda_{kl}^{p+1}|^2 \leqslant \max_{k,l}|\lambda_{kl}|^2\sum_{k,l}|c_{kl}\lambda_{kl}^{p}|^2 =$$
$$= \max_{k,l}|\lambda_{kl}|^2(u^{p}, u^{p}).$$

L'égalité a lieu si on s'est donné pour $\psi(mh, nh)$ la fonction propre $\psi_{mn}^{(k,l)}$ de l'opérateur $E+\tau(\Lambda_{xx}+\Lambda_{yy})$ de passage du niveau $t = p\tau$ au niveau $t = (p+1)\tau$ associée à la valeur propre λ_{kl} de plus grand module.

Si $\max\limits_{k,l} |\lambda_{kl}| \leqslant 1$, on a l'inégalité de l'énergie

$$(u^{p+1}, u^{p+1}) \leqslant (u^p, u^p). \tag{20}$$

Lorsque k et l prennent les valeurs $k, l = 1, 2, \ldots, M-1$, les valeurs propres parcourent un ensemble fini de points de la droite réelle à gauche du point $\lambda = 1$. Le point le plus à gauche s'obtient pour $k = l = M-1$:

$$\lambda_{M-1,\,M-1} = 1-8r\sin^2\frac{(M-1)\pi}{2M} = 1-8r\cos^2\frac{\pi}{2M} = 1-8r+O\left(\frac{1}{M^2}\right),$$

et l'inégalité $\max|\lambda_{kl}| \leqslant 1$ a donc lieu pour $-1 \leqslant 1-8r$, $r \leqslant 1/4$.

Dans le cas du schéma aux différences implicite

$$\left.\begin{aligned} \frac{u_{mn}^{p+1}-u_{mn}^{p}}{\tau} &= \Lambda_{xx}u_{mn}^{p+1}+\Lambda_{yy}u_{mn}^{p+1}, \\ u_{mn}^{0} &= \psi(mh, nh), \\ u_{mn}^{p}|_{\Gamma} &= 0 \end{aligned}\right\}$$

la solution prend la forme

$$u_{mn}^{p} = \sum_{k,l} c_{kl}\lambda_{kl}^{p}\psi_{mn}^{(k,l)},$$

où

$$\lambda_{kl} = \frac{1}{1+4r\left(\sin^2\frac{k\pi}{2M}+\sin^2\frac{l\pi}{2M}\right)},$$

les coefficients c_{kl} étant toujours définis par la formule (19). Ici $0 < \lambda_{kl} < 1$, et l'inégalité de l'énergie (20) a lieu quel que soit $r = \tau/h^2$.

4. Représentation de la solution d'un schéma aux différences pour le problème des cordes vibrantes. Prenons un schéma à trois niveaux $L_h u^{(h)} = f^{(h)}$ approchant le problème relatif aux vibrations d'une corde aux extrémités fixées:

$$\left.\begin{aligned} u_{tt}-u_{xx} &= 0, & 0 \leqslant x \leqslant 1, \quad 0 \leqslant t \leqslant T, \\ u(0, t) = u(1, t) &= 0, & 0 \leqslant t \leqslant T, \\ u(x, 0) &= \psi_0(x), & 0 \leqslant x \leqslant 1, \\ u_t(x, 0) &= \psi_1(x), & 0 \leqslant x \leqslant 1. \end{aligned}\right\}$$

Posons

$$L_h u^{(h)} \equiv \left\{\begin{aligned} &\frac{u_m^{p+1}-2u_m^{p}+u_m^{p-1}}{\tau^2}-\Lambda_{xx}u_m^{p} = 0, \\ &u_0^{p} = u_M^{p} = 0, \\ &u_m^{0} = \psi_0(mh), \\ &u_m^{1} = \tilde{\psi}_m, \end{aligned}\right.$$

où

$$\tilde{\varphi}_m = u(mh, 0)+\tau u_t(mh, 0)+\frac{\tau^2}{2} u_{tt}(mh, 0) = \varphi_0(mh)+\tau\varphi_1(mh)+\frac{\tau^2}{2}\varphi_0''(x).$$

Cherchons les solutions de l'équation aux différences qui vérifient les conditions $u_0^p = u_M^p = 0$ sous la forme

$$u_m^p = \lambda^p\sqrt{2}\sin\frac{k\pi m}{M}\ (\equiv \lambda^p\psi^{(k)}) \tag{21}$$

sans imposer pour le moment les conditions initiales $u_m^0 = \varphi_0(mh)$ et $u_m^1 = \tilde{\varphi}_m$.

On obtient l'équation suivante pour λ :

$$\frac{\lambda-2+\frac{1}{\lambda}}{\tau^2}-\mu_k = 0, \qquad \mu_k = -\frac{4}{h^2}\sin^2\frac{k\pi}{2M},$$

$$\lambda^2-2\left(1-2r^2\sin^2\frac{k\pi}{2M}\right)+1 = 0, \qquad r = \frac{\tau}{h},$$

$$\lambda_1(k) = 1-2r^2\sin^2\frac{k\pi}{2M}+\sqrt{\left(1-2r^2\sin^2\frac{k\pi}{2M}\right)^2-1},$$

$$\lambda_2(k) = 1-2r^2\sin^2\frac{k\pi}{2M}-\sqrt{\left(1-2r^2\sin^2\frac{k\pi}{2M}\right)^2-1}.$$

Il existe donc deux solutions de la forme cherchée (21) :

$$\lambda_1^p(k)\psi^{(k)} \quad \text{et} \quad \lambda_2^p(k)\psi^{(k)}.$$

Le problème étant linéaire, l'expression

$$u_m^p = \sum_{k=1}^{M-1}[\alpha_k\lambda_1^p(k)+\beta_k\lambda_2^p(k)]\psi_m^{(k)}$$

est solution pour α_k et β_k, $k = 1, 2, \ldots, M-1$, choisis arbitrairement. Pour $p = 0$ et $p = 1$, on a respectivement

$$u_m^0 = \varphi_0(mh) = \sum_{k=1}^{M-1}(\alpha_k+\beta_k)\psi^{(k)},$$

$$u_m^1 = \tilde{\varphi}_m = \sum_{k=1}^{M-1}[\alpha_k\lambda_1(k)+\beta_k\lambda_2(k)]\psi^{(k)}.$$

Ces relations définissent les valeurs de α_k et β_k. La somme $\alpha_k+\beta_k$ doit être le coefficient de Fourier du développement de $\varphi_0(mh)$ suivant les fonctions $\{\psi^{(k)}\}$, i.e.

$$\alpha_k+\beta_k = h\sum_{m=0}^{M}\varphi_0(mh)\left(\sqrt{2}\sin\frac{k\pi m}{M}\right).$$

De même

$$\alpha_k\lambda_1(k)+\beta_k\lambda_2(k) = h\sum_{m=0}^{M}\tilde{\varphi}_m\left(\sqrt{2}\sin\frac{k\pi m}{M}\right).$$

L'écriture de la solution d'une équation aux différences sous forme de série finie de Fourier ne permet pas que de mettre en évidence les conditions imposées aux inégalités de l'énergie. Dans la suite de notre exposé nous nous servirons largement de telles représentations lors de l'étude qualitative de problèmes modèles, et cela dans des buts différents.

Il faut dire que ces représentations ne sont d'un usage immédiat pour

le calcul de solutions que dans des cas rares. Le fait est que les méthodes numériques commodes doivent s'appliquer à une vaste classe de problèmes. Les schémas aux différences écrits ci-dessus se généralisent aisément au cas des coefficients variables et des domaines irréguliers, et nous pouvons espérer que des propriétés telles que l'inégalité de l'énergie se conservent. Or, une telle modification du problème rend impossible la représentation de sa solution sous forme de série de Fourier : il est généralement impossible d'écrire les fonctions propres de l'opérateur de passage d'un niveau à l'autre et de calculer les valeurs propres associées.

Exercices

1. Etant donné le problème plan de la chaleur dans un domaine carré de valeurs nulles sur la frontière, considérer pour lui le schéma aux différences

$$\left.\begin{aligned} \frac{u_{mn}^{p+1}-u_{mn}^{p}}{\tau} &= \sigma[\Lambda_{xx}u_{mn}^{p+1}+\Lambda_{yy}u_{mn}^{p+1}]+ \\ &+(1-\sigma)[\Lambda_{xx}u_{mn}^{p}+\Lambda_{yy}u_{mn}^{p}], \quad 0 < mh, \quad nh < 1, \\ u_{mn}^{p}|_{\Gamma} &= 0, \quad u_{mn}^{0} = \varphi(mh, nh) \end{aligned}\right\}$$

(pour les notations voir le texte du paragraphe) et écrire la solution du problème sous forme de série de Fourier. Pour quelles valeurs du paramètre σ, $0 \leqslant \sigma \leqslant 1$, on a l'inégalité de l'énergie $(u^{p+1}, u^{p+1}) \leqslant (u^{p}, u^{p})$ pour n'importe quel rapport des pas $r = \tau/h^2$?

Pour quels σ un $u^{p} \neq 0$ quelconque vérifie l'inégalité stricte $(u^{p+1}, u^{p+1}) < (u^{p}, u^{p})$, le choix de r et h étant arbitraire ?

2. Soient le problème différentiel

$$u_t - u_{xx} = 0, \quad 0 \leqslant x \leqslant 1, \quad 0 \leqslant t \leqslant T,$$
$$u|_{\Gamma} = 0, \quad u(x, 0) = \varphi(x)$$

et le problème aux différences

$$\left.\begin{aligned} &\frac{u_{mn}^{p+1}-u_{mn}^{p}}{\tau} - \Lambda_{xx}u_{mn}^{p} = 0, \\ &u_{mn}^{p}|_{\Gamma} = 0, \quad u_{mn}^{0} = \varphi(mh, nh). \end{aligned}\right\}$$

Mettre leurs solutions sous forme de série de Fourier et de série finie de Fourier respectivement. En comparant ces séries pour $r \leqslant {}^1/_2$, démontrer sous l'hypothèse de $\varphi''(x)$ bornée que la solution du problème aux différences converge vers celle du problème différentiel. Démontrer que pour $r > {}^1/_2$ la convergence n'a en général pas lieu.

3. Représenter par la série finie de Fourier la solution du problème de Dirichlet aux différences relatif à l'équation de Poisson dans le domaine carré $0 \leqslant x, y \leqslant 1$:

$$\Lambda_{xx}u_{mn}+\Lambda_{yy}u_{mn} = \varphi(mh, nh), \quad 0 < mh, nh < 1,$$

avec la condition aux limites

a) $$u_{mn}|_{\Gamma} = 0,$$

b) $$u_{mn} = \begin{cases} nh & \text{si } m = 0, \\ 1+mh & \text{si } n = M, \\ mh & \text{si } n = 0, \\ 1+nh & \text{si } m = M. \end{cases}$$

INDICATION pour 3, b): $u_{mn} = mh+nh+Z_{mn}$, Z_{mn} vérifiant les conditions homogènes sur la frontière.

§28. Principe du maximum

Nous avons vu plus haut (§§21 et 24) comment on démontre la stabilité moyennant le principe du maximum. Dans les deux exemples intéressants qui suivent, cet artifice le permet une fois de plus. Nous allons considérer des schémas aux différences explicite et implicite approchant le problème aux limites pour l'équation de la chaleur :

$$\left.\begin{aligned} \frac{\partial u}{\partial t}-a^2(x,t)\frac{\partial^2 u}{\partial x^2} &= \varphi(x,t), \quad 0<x<1, \quad 0\leqslant t\leqslant T,\\ u(x,0) &= \psi_0(x), \quad 0<x<1,\\ u(0,t) &= \psi_1(t), \quad 0\leqslant t\leqslant T,\\ u(1,t) &= \psi_2(t), \quad 0\leqslant t\leqslant T. \end{aligned}\right\} \tag{1}$$

1. Schéma aux différences explicite. Soit le schéma aux différences explicite

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{n+1}-u_m^n}{\tau}-a^2(mh,n\tau)\dfrac{u_{m+1}^n-2u_m^n+u_{m-1}^n}{h^2} = \varphi(mh,n\tau),\\ m=1,2,\ldots,M-1; \quad n=0,1,\ldots,[T/\tau]-1,\\ u_m^0=\psi_0(mh), \qquad m=0,1,\ldots,M,\\ u_0^n=\psi_1(n\tau), \qquad n=1,2,\ldots,[T/\tau],\\ u_M^n=\psi_2(n\tau), \qquad n=1,2,\ldots,[T/\tau]. \end{cases} \tag{2}$$

Ici $M = 1/h$ est un entier.

Le critère spectral de von Neumann réuni au principe des coefficients fixés conduit, nous l'avons constaté au §26, à une condition nécessaire de stabilité

$$\frac{\tau}{h^2} \leqslant \frac{1}{2\max\limits_{x,t} a^2(x,t)}. \tag{3}$$

Démontrons que dans cette condition la stabilité a effectivement lieu si les normes se définissent par les égalités

$$\left.\begin{aligned} \|u^{(h)}\|_{U_h} &= \max_n \max_m |u_m^n|,\\ \|f^{(h)}\|_{F_h} &= \max\Big(\max_m |\psi_0(mh)|, \max_n |\psi_1(n\tau)|,\\ &\qquad \max_n |\psi_2(n\tau)|, \max_{m,n} |q(x_m,t_n)|\Big). \end{aligned}\right\} \tag{4}$$

Etablissons la validité de l'inégalité *(principe du maximum)*

$$\max_m |u_m^{n+1}| \leqslant \max\Big[\max_n |\psi_1(t_n)|, \max_n |\psi_2(t_n)|, \max_m |u_m^n| + \tau \max_{m,n} |q(x_m,t_n)|\Big]. \tag{5}$$

A cette fin récrivons comme suit l'équation aux différences à la base du

schéma (2) :

$$u_m^{n+1} = (1-2ra^2(x_m, t_n))u_m^n + ra^2(x_m, t_n)(u_{m-1}^n + u_{m+1}^n) + \\ + \tau\varphi(x_m, t_n), \quad m = 1, \ldots, M-1. \quad (6)$$

Dans la condition (3) l'expression $1-2ra^2(x_m, t_n)$ est non négative et on écrit donc

$$|u_m^{n+1}| \leqslant [1-2ra^2(x_m, t_n)] \max_k |u_k^n| + ra^2(x_m, t_n)\Big(\max_k |u_k^n| + \\ + \max_k |u_k^n|\Big) + \tau \max_{m,n} |\varphi(x_m, t_n)| = \max_k |u_k^n| + \\ + \tau \max_{m,n} |\varphi(x_m, t_n)|, \quad m = 1, 2, \ldots, M-1. \quad (7)$$

Vu que

$$u_0^{n+1} = \psi_1[(n+1)\tau], \quad u_M^{n+1} = \psi_2[(n+1)\tau], \quad (8)$$

on en déduit le principe du maximum (5).

Mettons la solution $u^{(h)}$ du problème $L_h u^{(h)} = f^{(h)}$ sous forme de somme de deux termes : $u^{(h)} = v^{(h)} + w^{(h)}$, $v^{(h)}$ et $w^{(h)}$ étant définis comme solutions des problèmes

$$L_h v^{(h)} = \begin{cases} 0 \\ \psi_0(x_m), \\ \psi_1(t_n), \\ \psi_2(\tau_n) \end{cases} \quad \text{et} \quad L_h w^{(h)} = \begin{cases} \varphi(x_m, t_n), \\ 0, \\ 0, \\ 0. \end{cases} \quad (9)$$

En vertu de l'estimation (5)

$$\max_m |v_m^{n+1}| \leqslant \max\Big[\max_k |\psi_1(t_k)|, \max_k |\psi_2(t_k)|, \max_m |v_m^n|\Big],$$

$$\max_m |v_m^n| \leqslant \max\Big[\max_k |\psi_1(t_k)|, \max_k |\psi_2(t_k)|, \max_m |v_m^{n-1}|\Big],$$

$$\max_m |v_m^{n-1}| \leqslant \max\Big[\max_k |\psi_1(t_k)|, \max_k |\psi_2(t_k)|, \max_m |v_m^{n-2}|\Big],$$

. . . . . .

$$\max_m |v_m^1| = \max\Big[\max_k |\psi_1(t_k)|, \max_k |\psi_2(t_k)|, \max_m |\psi_0(x_m)|\Big].$$

Toujours en vertu de (5) on a

$$\max_m |w_m^{n+1}| \leqslant \max_m |w_m^n| + \tau \max_{m,k} |\varphi(x_m, t_k)| \leqslant \\ \leqslant \max_m |w_m^{n-1}| + 2\tau \max_{m,k} |\varphi(x_m, t_k)| \leqslant$$

. . . . . .

$$\leqslant \max_m |w_m^0| + (n+1)\tau \max_{m,k} |\varphi(x_m, t_k)| \leqslant T \max_{m,k} |\varphi(x_m, t_k)|.$$

Les estimations de v_m^{n+1} et w_m^{n+1} entraînent

$$\max_m |u_m^{n+1}| = \max_m |v_m^{n+1}+w_m^{n+1}| \leqslant \max_m |v_m^{n+1}|+\max_m |w_m^{n+1}| \leqslant$$
$$\leqslant \max\left[\max_k |\psi_1(t_k)|, \max_k |\psi_2(t_k)|, \max_m |\psi_0(x_m)|\right]+$$
$$+T\max_{m,n} |\varphi(x_m, t_n)| \leqslant c\,||f^{(h)}||_{F_h}, \tag{10}$$

où

$$c = 2\max(1, T). \tag{11}$$

L'inégalité est juste pour tous les n. Donc

$$|u^{(h)}||_{U_h} \leqslant c\,||f^{(h)}||_{F_h}, \tag{12}$$

et on a la stabilité.

2. Schéma aux différences implicite. Voyons maintenant le schéma implicite

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{n+1}-u_m^n}{\tau} - a^2(x_m, t_n)\dfrac{u_{m+1}^{n+1}-2u_m^{n+1}+u_{m-1}^{n+1}}{h^2} = \varphi(x_m, t_n), \\ u_m^0 = \psi_0(x_m), \\ u_0^n = \psi_1(t_n), \\ u_M^n = \psi_2(t_n). \end{cases} \tag{13}$$

Connaissant u_m^n, $m = 0, 1, \ldots, M$, on calcule les valeurs u_m^{n+1}, $m = 0, 1, \ldots, M$, en résolvant le problème

$$\left.\begin{aligned} \frac{u_m^{n+1}}{\tau} - a^2(x_m, t_n)\frac{u_{m+1}^{n+1}-2u_m^{n+1}+u_{m-1}^{n+1}}{h^2} = \frac{u_m^n}{\tau}+\varphi(x_m, t_n), \\ u_0^{n+1} = \psi_1(t_{n+1}), \quad u_M^{n+1} = \psi_2(t_{n+1}). \end{aligned}\right\} \tag{14}$$

Après avoir multiplié les deux membres de l'équation aux différences par $-\tau$ ce dernier s'écrit :

$$\left.\begin{aligned} a_m v_{m-1}+b_m v_m+c_m v_{m+1} = g_m, \quad m = 1, 2, \ldots, M-1, \\ v_0 = \alpha, \quad v_M = \beta, \end{aligned}\right\} \tag{15}$$

avec

$$v_m = u_m^{n+1}, \quad a_m = a^2(x_m, t_n)r,$$
$$b_m = -2a^2(x_m, t_n)r-1, \quad c_m = a^2(x_m, t_n)r,$$
$$g_m = -u_m^n-\tau\varphi(x_m, t_n), \quad \alpha = \psi_1(t_{n+1}), \quad \beta = \psi_2(t_{n+1}).$$

Les coefficients a_m, b_m, c_m satisfont aux conditions

$$a_m > 0, \quad c_m > 0, \quad |b_m| > a_m+c_m+\delta \quad (\delta > 0).$$

C'est pourquoi (voir §§4, 5), le problème admet une solution unique

$$(v_0, v_1, \ldots, v_M) = (u_0^{n+1}, u_1^{n+1}, \ldots, u_M^{n+1})$$

qui s'obtient par balayage.

Il nous reste à démontrer l'inégalité (12). A cette fin on montre qu'il y a (5) (principe du maximum) d'où les estimations (10) et (12) se déduisent tout comme dans le cas du schéma explicite (2).

De toutes les valeurs u_m^{n+1} de module égal à $\max\limits_m |u_m^{n+1}|$ prenons celle dont l'indice m prend la valeur minimale $m = m^*$. Si $m^* = 0$ ou $m^* = M$, l'inégalité (5) est évidente conformément à (8). Soient $m^* \neq 0$ et $m^* \neq M$. Ecrivons l'équation (14) correspondant à cette valeur $m = m^*$:

$$ra^2(x_{m^*}, t_n)u_{m^*-1}^{n+1} - (1 + 2ra^2(x_{m^*}, t_n))u_{m^*}^{n+1} + \\ + ra^2(x_{m^*}, t_n)u_{m^*+1}^{n+1} = -u_{m^*}^n - \tau\varphi(x_{m^*}, t_n).$$

Supposons, pour fixer les idées, que $u_{m^*}^{n+1} > 0$. Le premier membre de cette égalité est alors évalué par

$$ra^2(x_{m^*}, t_n)u_{m^*-1}^{n+1} - (1 + 2ra^2(x_{m^*}, t_n))u_{m^*}^{n+1} + ra^2(x_{m^*}, t_n)u_{m^*+1}^{n+1} = \\ = a^2(x_{m^*}, t_n)\,[(u_{m^*-1}^{n+1} - u_{m^*}^{n+1}) + (u_{m^*+1}^{n+1} - u_{m^*}^{n+1})] - u_{m^*}^{n+1} \leqslant -u_{m^*}^{n+1}.$$

Donc

$$-u_{m^*}^{n+1} \geqslant -u_{m^*}^n - \tau\varphi(x_{m^*}, t_n),$$

$$\max_m |u_m^{n+1}| = u_{m^*}^{n+1} \leqslant |u_{m^*}^n + \tau\varphi(x_{m^*}, t_n)| \leqslant \\ \leqslant \max_m |u_m^n| + \tau \max_{m,n} |\varphi(x_m, t_n)|.$$

3. Comparaison des schémas aux différences explicite et implicite. Ainsi, nous avons démontré l'inégalité (5) qui signifie le principe du maximum et, simultanément, la stabilité du schéma aux différences implicite (14) dans les normes (4).

Soulignons la différence essentielle entre le schéma explicite (2) *et celui implicite* (14). *Le premier est stable à condition d'imposer la contrainte au pas*

$$\tau \leqslant \frac{1}{2 \max a^2(x, t)} h^2$$

qui devient très sévère si le coefficient $a^2(x, t)$ prend de grandes valeurs dans un voisinage, fût-il restreint, d'un seul point, tandis que le second reste stable pour n'importe quel rapport des pas h et τ.

On dit des schémas aux différences qui restent, comme le schéma implicite (14), stables pour un rapport quelconque des pas du réseau qu'ils sont *absolument* (ou *inconditionnellement) stables*. Ce n'est pas le cas du schéma explicite (2).

CHAPITRE 9

LES SCHÉMAS AUX DIFFÉRENCES ET LE CALCUL DES SOLUTIONS GÉNÉRALISÉES

§29. Solution généralisée

Dans tous les exemples ci-dessus, nous avons supposé l'existence de solutions suffisamment régulières des problèmes aux limites différentiels, et la construction de schémas aux différences se faisait en remplaçant de façon approchée les dérivées de l'équation différentielle par des rapports aux différences. De nombreux processus physiques importants sont cependant régis par des fonctions autres que les fonctions dérivables. Ainsi, les expériences physiques montrent que la répartition de pression, de densité et de température dans l'écoulement supersonique d'un gaz non visqueux est décrite par des fonctions présentant des sauts (des ondes de choc). Les sauts peuvent se produire avec le temps pour des données initiales régulières.

Les problèmes aux limites différentiels correspondants ne possédant pas de solutions régulières, il faut élargir la notion de solution et introduire d'une façon naturelle des solutions généralisées qui peuvent être discontinues. Décrivons deux procédés principaux dont on se sert en l'occurrence.

Le premier consiste à écrire les lois physiques de conservation (de la masse, de l'impulsion, de l'énergie, etc.) sous forme intégrale au lieu de la forme différentielle. Les lois ont alors un sens pour des fonctions discontinues qui se prêtent à une intégration mais qu'il est impossible de dériver.

Dans le second procédé on introduit artificiellement dans les équations différentielles des termes tels que celles-ci aient des solutions régulières. S'agissant des problèmes relatifs à la mécanique des fluides, ces termes ont le sens d'une petite viscosité lissant les discontinuités de l'écoulement. On fait ensuite tendre vers zéro les coefficients des termes « de viscosité » et on prend la limite de la solution pour solution généralisée du problème originel.

Nous éclaircirons la notion de solution généralisée et les procédés de son calcul à l'aide du problème de Cauchy suivant :

$$\left.\begin{aligned} \frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x} &= 0, \qquad 0<t<T,\ -\infty<x<\infty, \\ u(x,0) &= \psi(x), \qquad\qquad -\infty<x<\infty, \end{aligned}\right\} \tag{1}$$

le plus simple modèle de problèmes de la mécanique des fluides, qui possèdent des solutions discontinues pour des données initiales régulières.

1. Apparition des discontinuités. Supposons d'abord que le problème (1) admet une solution $u(x, t)$ régulière et introduisons des lignes $x = x(t)$ définies par l'équation

$$\frac{dx}{dt} = u(x, t). \tag{2}$$

Ces lignes s'appellent *caractéristiques* de l'équation $u_t + uu_x = 0$. Le long de chaque caractéristique $x = x(t)$ la solution $u(x, t)$ peut être assimilée à une fonction dépendant de t seul :

$$u(x, t) = u[x(t), t] = u(t).$$

Alors

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} = \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0.$$

La solution est donc constante le long de la caractéristique : $u(x, t) = \text{const}$. Mais en vertu de l'équation (2), il résulte de $u = \text{const}$ que la caractéristique est une droite $x = ut + x_0$. Ici x_0 est l'abscisse du point $(x_0, 0)$, origine de la caractéristique, et $u = \psi(x_0)$ le coefficient angulaire. Il suffit donc de définir la fonction de départ $u(x, 0) = \psi(x)$ pour avoir une représentation sensible des caractéristiques et des valeurs de la solution $u(x, t)$ en chaque point du demi-plan $t > 0$ (fig. 28).

Notons que sous l'hypothèse de $u(x, t)$ régulière les caractéristiques ne peuvent pas se couper, parce que, dans le cas contraire, chacune d'elles

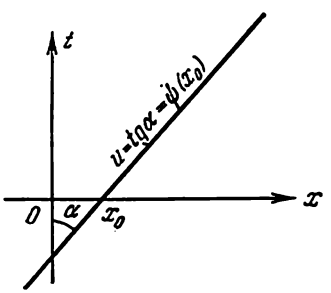


Fig. 28

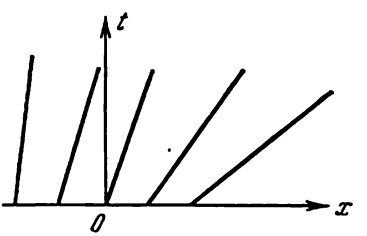


Fig. 29

apporterait au point d'intersection sa valeur de la solution à elle et la solution ne serait donc pas une fonction univoque. Pour $\psi(x)$ croissant de façon monotone, l'angle α croît avec x_0 et les caractéristiques ne se rencontrent pas (fig. 29). Mais si la fonction $\psi(x)$ décroît, les caractéristiques convergent et se coupent donc inévitablement que la fonction $\psi(x)$ soit régulière ou non. Le problème (1) n'admet pas de solution régulière dès l'instant $t = \bar{t}$ lorsque au moins deux caractéristiques se coupent (fig. 30).

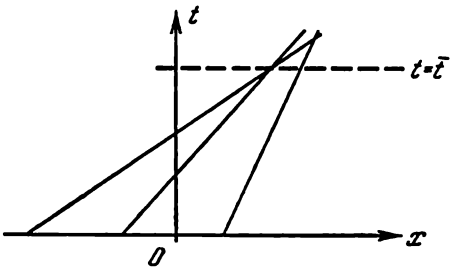


Fig. 30

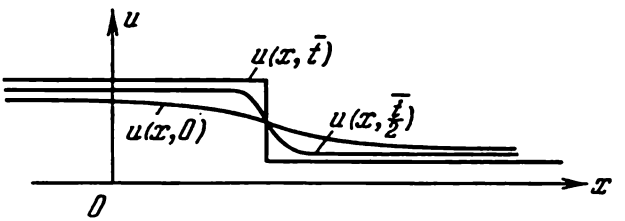


Fig. 31

Les graphes de la fonction $u = u(x, t)$ pour $t = 0$, $^1/_2\,\bar{t}$ et $\bar{t}$ sont représentés sur la figure 31.

2. Définition d'une solution généralisée. Rappelons la formule de Green qui nous permettra de définir la solution généralisée du problème (1). Soit D un domaine de frontière Γ dans le plan Oxt et soient $\Phi_1(x, t)$ et $\Phi_2(x, t)$ deux fonctions possédant dans D des dérivées partielles continues jusqu'à sur Γ. On a alors la formule de Green :

$$\iint\limits_D \left(\frac{\partial \Phi_1}{\partial t} + \frac{\partial \Phi_2}{\partial x}\right) dx\, dt = \oint\limits_\Gamma \Phi_1 dx - \Phi_2 dt. \tag{3}$$

$\dfrac{\partial \Phi_1}{\partial t} + \dfrac{\partial \Phi_2}{\partial x}$ est la *divergence* du vecteur $\Phi = \begin{pmatrix}\Phi_1\\ \Phi_2\end{pmatrix}$. La formule de Green (3) signifie que l'intégrale de la divergence du champ vectoriel Φ étendue au domaine D est égale au flux du vecteur Φ à travers la frontière Γ de ce domaine.

Ecrivons l'équation différentielle du problème (1) sous la forme d'une divergence :

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0 \tag{4}$$

et intégrons les deux membres de (4) sur un domaine D quelconque du demi-plan $t \geqslant 0$, il vient

$$0 = \iint\limits_D \left[\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)\right] dx\, dt = \oint\limits_\Gamma u\, dx - \frac{u^2}{2}\, dt.$$

Ainsi, chaque solution dérivable de (4) vérifie la relation intégrale

$$\oint_{\Gamma} u\,dx - \frac{u^2}{2}\,dt = 0, \tag{5}$$

Γ étant un contour du demi-plan $t > 0$. L'égalité (5) exprime une loi de conservation : le flux du vecteur $\begin{pmatrix} u \\ u^2/2 \end{pmatrix}$ à travers tout contour fermé est nul.

Montrons que, inversement, si une fonction régulière satisfait pour tout contour Γ à la loi intégrale de conservation (5), en chaque point (x_0, t_0), $t_0 > 0$, on a l'égalité (4). Raisonnons par l'absurde et supposons, pour fixer les idées, qu'en un point (x_0, t_0) on a

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)\Bigg|_{\substack{x=x_0\\ t=t_0}} > 0.$$

Vu la continuité, il existe alors un cercle D de circonférence Γ et de centre le point (x_0, t_0) tellement petit qu'en tout son point $u_t + \left(\frac{u^2}{2}\right)_x > 0$. On obtient

$$0 = \oint_{\Gamma} u\,dx - \frac{u^2}{2}\,dt = \iint_D \left[\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)\right] dx\,dt > 0.$$

La contradiction $0 > 0$ montre que (5) entraîne (4) pour $u(x, t)$ régulier de sorte que les deux égalités sont équivalentes. Si, par contre, $u(x, t)$ est discontinu, l'équation différentielle (1) ou (4) n'a plus de sens sur la ligne de discontinuité mais la condition intégrale (5) a toujours un sens. Aussi appellerons-nous *solution généralisée* de l'équation (4) toute fonction dérivable par morceaux qui vérifie, quel que soit le contour Γ choisi dans le demi-plan $t \geqslant 0$, la condition (5).

3. Condition sur la ligne de discontinuité de la solution. Supposons qu'on cherche la solution dans un domaine contenant une ligne $x \equiv x(t)$ sur laquelle la solution généralisée $u(x, t)$ subit une discontinuité de première espèce. En s'approchant de cette ligne à gauche et à droite on obtient respectivement

$$u(x, t) = u_{\text{gauche}}(x, t),$$
$$u(x, t) = u_{\text{droite}}(x, t).$$

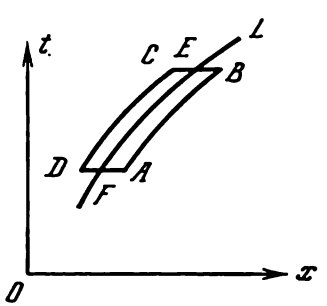


Fig. 32

Il se trouve que les valeurs $u_{\text{gauche}}(x, t)$, $u_{\text{droite}}(x, t)$ et la vitesse $\dot{x} = dx/dt$ du point de discontinuité ne peuvent être quelconques : il y a entre elles une certaine relation.

Soit L la ligne de discontinuité (fig. 32). L'intégrale $\int_{ABCDA} u\,dx - \frac{u^2}{2}\,dt$ sur le contour

ABCDA ou sur tout autre contour s'annule. Les segments *BC* et *DA* se réduisent alors aux points *E* et *F* respectivement et les intégrales correspondantes s'annulent. On a

$$\int_{L'} [u]\,dx - \left[\frac{u^2}{2}\right] dt = 0,$$

ou

$$\int_{L'} \left([u]\,\frac{dx}{dt} - \left[\frac{u^2}{2}\right]\right) dt = 0,$$

où $[z] = z_{\text{droite}} - z_{\text{gauche}}$ est le saut de la quantité z sur la ligne de discontinuité et L' un tronçon arbitraire de celle-ci : $L' = EF$.

Le caractère arbitraire de L' fait qu'en tout point de L la fonction à intégrer s'annule :

$$[u]\,\frac{dx}{dt} - \left[\frac{u^2}{2}\right] = 0.$$

D'où

$$\frac{dx}{dt} = \left[\frac{u^2}{2}\right] : [u] = \frac{u^2_{\text{droite}} - u^2_{\text{gauche}}}{2(u_{\text{droite}} - u_{\text{gauche}})} = \frac{u_{\text{gauche}} + u_{\text{droite}}}{2},$$

de sorte que la ligne de discontinuité vérifie la condition

$$\frac{dx}{dt} = \frac{u_{\text{gauche}} + u_{\text{droite}}}{2}. \tag{6}$$

En mettant l'équation $u_t + uu_x = 0$ sous la forme d'une autre divergence, à savoir

$$\frac{\partial}{\partial t}\left(\frac{u^2}{2}\right) + \frac{\partial}{\partial x}\left(\frac{u^3}{3}\right) = 0, \tag{7}$$

nous aurions abouti de façon analogue à une autre relation intégrale

$$\oint_{\Gamma} \frac{u^2}{2}\,dx - \frac{u^3}{3}\,dt = 0 \tag{8}$$

et à une autre condition sur la ligne de discontinuité

$$\frac{dx}{dt} = \frac{2}{3}\,\frac{u^2_{\text{gauche}} + u_{\text{gauche}}u_{\text{droite}} + u^2_{\text{droite}}}{u_{\text{gauche}} + u_{\text{droite}}}. \tag{9}$$

La pente (9) de la ligne de discontinuité (ou la vitesse $\dot{x}$ de l'onde de choc) ne coïncide pas avec la pente (6) correspondant à la forme (4). On voit donc que la notion de solution généralisée dépend de la loi intégrale de conservation traduite par l'équation différentielle donnée (1). Dans les problèmes de physique mathématique les lois intégrales de conservation ont un sens physique précis.

Pour les fonctions u régulières les cinq écritures

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}=0,$$

$$\frac{\partial u}{\partial t}+\frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)=0,$$

$$\frac{\partial}{\partial t}\left(\frac{u^2}{2}\right)+\frac{\partial}{\partial x}\left(\frac{u^3}{3}\right)=0,$$

$$\oint_\Gamma u\,dx-\frac{u^2}{2}\,dt=0,$$

$$\oint_\Gamma \frac{u^2}{2}\,dx-\frac{u^3}{3}\,dt=0$$

sont équivalentes.

Lorsque nous aborderons le problème de Cauchy (1), nous supposerons remplies la loi intégrale de conservation (5) et la condition (6) sur la ligne de discontinuité qui en découle.

4. Suppression d'une discontinuité. Etant donné les données initiales discontinues

$$u=\begin{cases}2 & \text{pour} \quad x<0,\\ 1 & \text{pour} \quad x>0,\end{cases}$$

la solution correspondante est visualisée sur la figure 33.

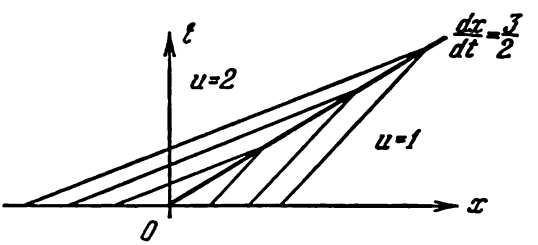


Fig. 33

La pente de la ligne de discontinuité $\frac{dx}{dt}=\frac{2+1}{2}=\frac{3}{2}$ est la moyenne arithmétique des pentes des caractéristiques de part et d'autre de la ligne.

Donnons-nous maintenant une autre discontinuité avec les conditions initiales :

$$u=\begin{cases}1 & \text{pour} \quad x<0,\\ 2 & \text{pour} \quad x>0.\end{cases}$$

La figure 34 montre deux méthodes de construction de la solution, dont l'une donne une solution continue et l'autre une solution discontinue pour $t>0$. Notre choix doit porter sur la solution continue et cela pour la raison

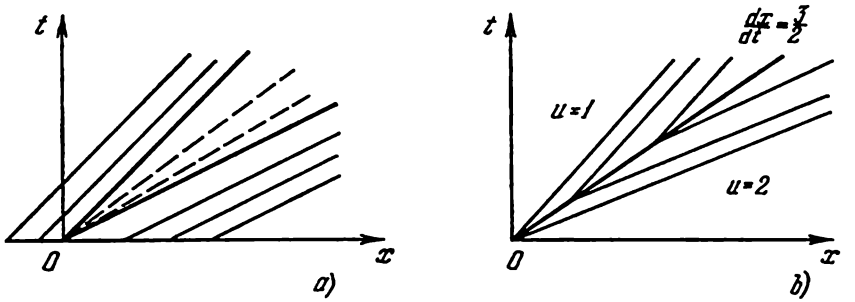


Fig. 34

suivante. En modifiant quelque peu les données initiales à force de les définir par la formule

$$u = \begin{cases} 1 & \text{pour} \quad x \leqslant 0, \\ 2 & \text{pour} \quad x \geqslant \varepsilon, \\ 1+x/\varepsilon & \text{pour} \quad 0 \leqslant x \leqslant \varepsilon, \end{cases}$$

on obtiendra une solution u définie de façon unique. Cette solution représentée sur la figure 35 devient, pour ε tendant vers zéro, la solution continue

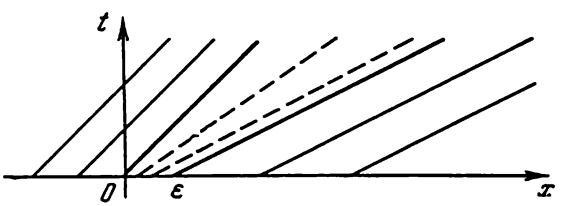


Fig. 35

de la figure 34, *a*. Le rejet de la solution de la figure 34, *b* à cause de son instabilité par rapport aux perturbations des données initiales est analogue au refus d'utiliser les ondes de choc de raréfaction dans la description mathématique de l'écoulement d'un gaz parfait.

5. Autre définition d'une solution généralisée. Pour définir ce qu'est la solution généralisée du problème (1) on peut considérer le problème auxiliaire

$$\left.\begin{aligned} \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} &= \mu\frac{\partial^2 u}{\partial x^2}, \\ u(x, 0) &= \psi(x). \end{aligned}\right\} \tag{10}$$

L'équation différentielle y figurant est du type parabolique. Ses solutions restent régulières si $\psi(x)$ est une fonction régulière. Dans le cas où $u(x, 0) = \psi(x)$ est discontinu, la discontinuité est lissée. Le paramètre $\mu > 0$ joue le rôle de viscosité de la mécanique des fluides. Lorsque $\mu \to 0$. la solution de

(10) tend vers une limite qu'on peut prendre pour solution généralisée du problème (1). On montre que s'agissant du problème (1) la dernière définition de la solution généralisée équivaut à celle utilisant la loi de conservation (5).

§30. Construction de schémas aux différences

Proposons-nous maintenant de construire des schémas aux différences pour le problème

$$\left.\begin{aligned} \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} &= 0, \\ u(x, 0) &= \psi(x). \end{aligned}\right\} \tag{1}$$

Supposons, pour fixer les idées, $\psi(x) > 0$. Alors $u(x, t) > 0$. Il est naturel d'aborder en premier lieu le schéma aux différences

$$\left.\begin{aligned} \frac{u_m^{p+1} - u_m^p}{\tau} + u_m^p \frac{u_m^p - u_{m-1}^p}{h} &= 0, \quad p = 0, 1, \ldots, \\ & \qquad\qquad m = 0, \pm 1, \ldots, \\ u_m^0 &= \psi(x_m). \end{aligned}\right\} \tag{2}$$

En fixant le coefficient u_m^p au point $m = m_0$, on constate que l'équation à coefficients constants qui apparaît vérifie lors du passage au niveau $t = (p+1)\tau$ le principe du maximum si le pas $\tau = \tau_p$ est choisi à partir de la condition

$$r_p = \frac{\tau_p}{h} \leqslant \frac{1}{\max\limits_m |u_m^p|}.$$

On peut donc s'attendre à la stabilité. Si la solution du problème (1) est régulière, celui-ci est évidemment approché par le problème (2). En effet, le calcul expérimental des solutions régulières données à priori confirme la convergence.

Mais dans le cas où le problème (1) admet une solution discontinue, on n'a aucun sens raisonnable à s'attendre à la convergence vers une solution généralisée donnée par exemple par la loi intégrale de conservation

$$\oint_\Gamma u\,dx - \frac{u^2}{2}\,dt = 0. \tag{3}$$

Car le schéma aux différences utilisé (2) ne comporte pas d'information sur la loi de conservation à la base de la définition de la solution généralisée. Or, cette loi peut être (8) du § 29, (3) ou une certaine autre.

En formant le schéma aux différences on utilisera donc soit une loi intégrale de conservation correspondant à la solution généralisée cherchée, disons la loi (3), soit l'équation avec viscosité artificielle (10) du §29 :

$$u_t + uu_x = \mu u_{xx} \tag{4}$$

qui sélectionne pour $\mu \to 0$ la solution généralisée qui nous intéresse.

1. Schéma avec viscosité artificielle. Commençons par dire que le schéma aux différences comportant une petite viscosité introduite artificiellement

$$\frac{u_m^{p+1}-u_m^p}{\tau}+u_m^p\,\frac{u_m^p-u_{m-1}^p}{h}=\mu\,\frac{u_{m-1}^p-2u_m^p+u_{m+1}^p}{h^2},$$

$$u_m^0=\psi(x_m)$$

admet une solution $u^{(h)}=\{u_m^p\}$ qui converge uniformément pour $h\to 0$ et un $\tau=\tau(h,\mu)$ suffisamment faible vers la solution généralisée cherchée du problème (1) en dehors des voisinages aussi petits qu'on le veut des lignes de discontinuité de celle-ci. Ceci étant, $\mu=\mu(h)$ doit tendre vers zéro de façon suffisamment lente pour $h\to 0$. Les calculs en mécanique des fluides utilisent avec profit divers schémas aux différences avec viscosité artificielle. L'inconvénient de ces derniers est l'étalement des discontinuités.

Arrêtons-nous plus en détail sur des schémas aux différences construits à l'aide de la loi intégrale de conservation (3).

Deux approches sont possibles. Dans l'une, on se sert, en plus de la loi (3), de la condition qu'elle entraîne

$$\frac{dx}{dt}=\frac{u_{\text{gauche}}+u_{\text{droite}}}{2} \tag{5}$$

et qui est remplie sur la ligne de discontinuité. Dans l'autre, les discontinuités ne sont pas mises en évidence et les formules de calcul sont les mêmes pour tous les points.

2. Méthode des caractéristiques. L'idée consistant à isoler la discontinuité dans le calcul de la solution généralisée est le mieux réalisée par la méthode des caractéristiques, une variante de la méthode des différences finies. L'évolution des discontinuités se faisant jour lors des calculs, i.e. avec l'augmentation de t, se calcule par des formules spéciales utilisant le rapport (5). En dehors des discontinuités, toutes les données ci-dessus de l'équation différentielle sont équivalentes. Aussi, en construisant les formules de calcul aux points des domaines de régularité, part-on de la loi de conservation écrite sous forme différentielle, i.e. de l'équation différentielle

$$\frac{\partial u}{\partial t}+u\,\frac{\partial u}{\partial x}=0.$$

Voici le schéma d'une version de la méthode des caractéristiques relative à notre cas. Fixons sur l'axe Ox les points $x_m=mh$ et supposons, pour préciser les idées, que la condition initiale $u(x,0)=\psi(x)$ est donnée par une fonction $\psi(x)$ régulière. Faisons partir de chaque point $(x_m,0)$ une caractéristique de l'équation $u_t+uu_x=0$.

Admettons, pour ne pas alourdir l'exposé, que, $\psi(x)$ étant donné, on peut choisir un τ tellement petit que sur tout segment temporel de longueur τ chaque caractéristique coupe au plus une caractéristique voisine.

Prenons de tels τ et traçons les droites $t=t_p=p\tau$. Considérons les points en lesquels les caractéristiques issues des $(x_m,0)$ coupent la droite $t=\tau$ et transportons-y suivant les caractéristiques les valeurs $u(x_m,0)=$ $=\psi(x_m)$ de la solution.

Si sur le segment $0 \leqslant t \leqslant \tau$ toutes les caractéristiques sont disjointes deux à deux, faisons le pas suivant : prolongeons les caractéristiques jusqu'à ce qu'elles coupent la droite $t = 2\tau$ et transportons suivant elles aux points d'intersection les valeurs de la solution. Si les caractéristiques ne se coupent pas durant la période $\tau < t \leqslant 2\tau$, faisons le pas suivant, et ainsi de suite jusqu'à ce qu'un segment $t_p < t < t_{p+1}$ soit le lieu de rencontre de deux caractéristiques, par exemple de celles qui partent des points $(x_m, 0)$ et $(x_{m+1}, 0)$ (fig. 36). Le milieu du segment $Q_{m+1}^{p+1}Q_m^{p+1}$ est alors pris pour origine de la discontinuité. Remplaçons Q_m^{p+1}, Q_{m+1}^{p+1} par un seul point Q en lui attribuant deux valeurs de la solution $u_{\text{gauche}} = u(Q_m^p)$ et $u_{\text{droite}} = u(Q_{m+1}^p)$. Du point Q menons la ligne de discontinuité jusqu'à son intersection avec la droite $t = t_{p+2}$. Le coefficient angulaire de la discontinuité est défini par la condition sur la ligne de discontinuité

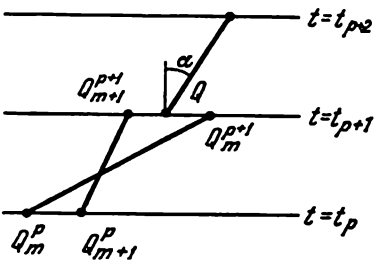


Fig. 36

$$\operatorname{tg}\alpha = \frac{u_{\text{gauche}} + u_{\text{droite}}}{2}.$$

Du point de rencontre de la discontinuité et de la droite $t = t_{p+2}$ menons les caractéristiques en arrière jusqu'à ce qu'elles coupent la droite $t = t_{p+1}$ tout en conservant les coefficients angulaires u_{gauche} et u_{droite} du niveau précédent. Par interpolation sur x trouvons aux points d'intersection de ces caractéristiques avec la droite $t = t_{p+1}$ les valeurs de u et prenons-les pour la valeur à gauche et celle à droite de la solution au point de discontinuité de la droite t_{p+2}. Cela permet de définir une nouvelle pente de la ligne de discontinuité comme la moyenne arithmétique des valeurs à gauche et à droite trouvées et de prolonger la discontinuité d'un pas en temps τ.

Le côté fort de la méthode des caractéristiques est qu'elle permet de suivre de près les discontinuités et de les calculer avec soin. En cours de calcul on voit pourtant surgir des discontinuités toujours nouvelles (en particulier, des discontinuités peu importantes peuvent se couper) si bien que la situation se complique avec le temps. La logique du calcul perd de sa simplicité tandis qu'augmentent l'encombrement de mémoire et le coût en temps machine.

C'est là l'inconvénient de la méthode des caractéristiques dans laquelle on isole les discontinuités et on les calcule de façon non standard.

3. Schémas aux différences conservatifs. Les schémas aux différences n'utilisant pas la viscosité artificielle ni les conditions sur la ligne de discontinuité s'appuient, nous l'avons constaté, sur des lois intégrales de conservation.

Construisons dans le plan Oxt le réseau des droites $t = p\tau$, $x = (m + 1/2)h$, $p = 0, 1, \ldots$; $m = 0, \pm 1, \ldots$, fixons les milieux des côtés des rectangles

ainsi formés (fig. 37) et rapportons-les au réseau D_h (les axes de coordonnées sont absents sur la figure).

La fonction cherchée $[u]_h$ sera assimilée à une fonction définie aux points de D_h en prenant la moyenne de la solution $u(x, t)$ sur le côté du rectangle élémentaire contenant le point considéré :

$$[u]_h\Big|_{\substack{x=x_m,\\ t=t_p}} \equiv \bar{u}_m^p = \frac{1}{h}\int\limits_{x_{m-1/2}}^{x_{m+1/2}} u(x, t_p)\,dx,$$

$$[u]_h\Big|_{\substack{x=x_{m+1/2},\\ t=t_{p+1/2}}} \equiv U_{m+1/2}^{p+1/2} = \frac{1}{\tau}\int\limits_{t_p}^{t_{p+1}} u(x_{m+1/2}, t)\,dt.$$

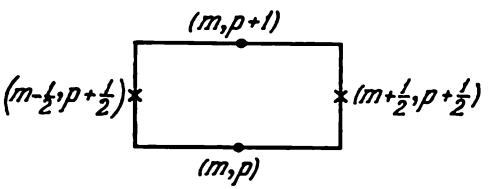


Fig. 37

La solution approchée $u^{(h)}$ du problème est définie sur le même réseau D_h. Désignons par u_m^p les valeurs de $u^{(h)}$ aux points (x_m, t_p) sur les côtés horizontaux des rectangles et par $U_{m+1/2}^{p+1/2}$ celles aux points $(x_{m+1/2}, t_{p+1/2})$ des côtés verticaux.

La quantité u_m^p peut être supposée prolongée sur tout le côté $t = t_p$, $x_m < x < x_{m+1}$ du rectangle dont fait partie le point (x_m, t_p). On admet de même que $U_{m+1/2}^{p+1/2}$ est défini sur tout le segment vertical

$$x = x_{m+1/2}, \quad t_p < t < t_{p+1}.$$

$u^{(h)}$ est donc défini sur les droites $x = mh$, $t = p\tau$. La relation entre u_m^p et $U_{m+1/2}^{p+1/2}$, $p = 0, 1, \ldots$; $m = 0, \pm 1, \ldots$, s'établit à partir de la loi intégrale de conservation

$$\oint\limits_{\Gamma} u\,dx - \frac{u^2}{2}\,dt = 0.$$

Soit Γ un rectangle élémentaire du réseau. Posons

$$-\oint\limits_{\Gamma} u^{(h)}\,dx - \frac{(u^{(h)})^2}{2}\,dt = 0 \tag{6}$$

ou en détaillant

$$h[u_m^{p+1} - u_m^p] + \frac{\tau}{2}\left[(U_{m+1/2}^{p+1/2})^2 - (U_{m-1/2}^{p+1/2})^2\right] = 0. \tag{7}$$

Connaissant la règle de calcul des quantités $U_{m+1/2}^{p+1/2}$, $m = 0, \pm 1, \ldots$, d'après les quantités connues u_m^p, $m = 0, \pm 1, \ldots$, la formule (7) permet de trouver u_m^{p+1}, $m = 0, \pm 1, \ldots$, i.e. d'avancer d'un pas temporel. Quel que soit le procédé choisi en vue d'obtenir $U_{m+1/2}^{p+1/2}$, un schéma aux différences de la forme (7) *traduit* cependant *des lois de conservation*. Voyons en quoi consiste cette propriété. Menons dans le demi-plan $t \geqslant 0$ un contour fermé sans point double formé de côtés de rectangles d'un réseau (fig. 38). Ce contour g_h limite un domaine G_h, réseau à mailles rectangulaires.

Sommons terme à terme toutes les équations (7) relatives aux rectangles de G_h.

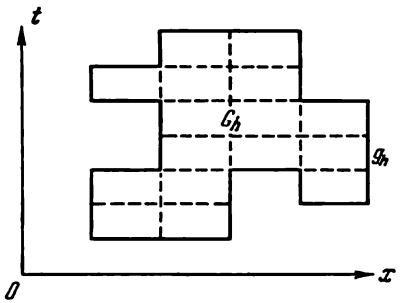


Fig. 38

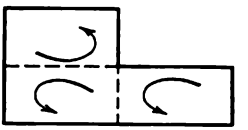


Fig. 39

Les équations (7) et (6) ne différant que par leur écriture, on peut estimer que nous sommons les équations (6). Il vient

$$-\oint_{g_h} u^{(h)}\,dx - \frac{(u^{(h)})^2}{2}\,dt = 0. \tag{8}$$

Les intégrales sur les côtés des rectangles intérieurs au contour g_h du domaine G_h mais figurant dans (6) se détruisent mutuellement après sommation des équations (6).

En effet, chacun de ces côtés appartient à deux rectangles voisins si bien que la fonction $u^{(h)}$ est intégrée sur eux deux fois et dans le sens contraire (fig. 39).

On appelle *conservatifs* des schémas aux différences tels qu'en les sommant sur les points du réseau G_h ne subsistent que les sommes algébriques des valeurs des inconnues ou des fonctions de ces inconnues le long de la frontière. Ces schémas sont analogues à des équations différentielles sous la forme d'une divergence

$$\operatorname{div}\Phi = \frac{\partial\Phi_1}{\partial t} + \frac{\partial\Phi_2}{\partial x} = 0$$

telles que leur intégration terme à terme sur un domaine bidimensionnel D fait apparaître dans le premier membre l'intégrale de contour (3) du §29. Le schéma aux différences (2) est non conservatif et le schéma (7) est conservatif.

Faisons une remarque. Supposons que la fonction $u^{(h)}$ vérifiant (7) converge uniformément pour $h \to 0$ vers une fonction continue par morceaux $u(x, t)$ dans tout domaine fermé ne contenant pas de lignes de discontinuité,

et soit $u^{(h)}$ bornée uniformément en h. Alors $u(x, t)$ satisfait à la loi intégrale de conservation

$$\oint_g u\,dx - \frac{u^2}{2}\,dt = 0,$$

avec g un contour régulier par morceaux arbitraire.

Cela découle directement de la possibilité d'approcher g par le contour g_h, de l'égalité (8) et de la convergence supposée(*) $u^{(h)} \to u$.

Pour que le schéma (7) ait un sens, il faut indiquer le procédé de calcul de $U_m^{p+1/2}$ en fonction de $u_{m+1/2}^p$. Le schéma de Godounov qui nous aide à illustrer la notion de schéma conservatif utilise à cet effet la solution du problème suivant sur la « suppression de la discontinuité ». Supposons qu'à l'instant initial la solution $u(x, 0)$ est donnée par les conditions

$$u(x, 0) = \begin{cases} u_{\text{gauche}} & \text{pour} \quad x < 0, \\ u_{\text{droite}} & \text{pour} \quad x > 0, \end{cases}$$

où $u_{\text{gauche}} = \text{const}$, $u_{\text{droite}} = \text{const}$. Il existe alors une solution généralisée correspondante que nous avons recherchée au §29 dans les exemples $u_{\text{gauche}} = 1$, $u_{\text{droite}} = 2$ et $u_{\text{gauche}} = 2$, $u_{\text{droite}} = 1$. Il nous faut connaître la valeur $U = u(0, t)$ de la solution $u(x, t)$ pour $x = 0$.

A force de construire des figures du type 33, 34 représentant $u(x, t)$, le lecteur vérifiera sans peine que sur la droite $x = 0$ la solution prend les valeurs u_{gauche}, u_{droite} ou 0 selon les données initiales et établira la valeur pour tout couple de nombres u_{gauche} et u_{droite}. Lorsque $u_{\text{gauche}} > 0$, $u_{\text{droite}} > 0$, on aura par exemple $u(0, t) \equiv u_{\text{gauche}}$, et la réponse sera $u(0, t) \equiv u_{\text{droite}}$ pour $u_{\text{gauche}} < 0$, $u_{\text{droite}} < 0$.

La quantité $U_{m+1/2}^{p+1/2} (= U)$ du schéma (7) sera définie moyennant le problème de la suppression de la discontinuité apparaissant sur la frontière $x = x_{m+1/2}$ de n'importe quels deux secteurs où sont données les valeurs constantes $u_m^p (= u_{\text{gauche}})$ et $u_{m+1}^p (= u_{\text{droite}})$.

Si par exemple $u_m^p > 0$, $m = 0, \pm 1, \ldots$, alors

$$U_{m+1/2}^{p+1/2} = u_{\text{gauche}} = u_m^p, \quad m = 0, \pm 1, \ldots,$$

et le schéma (7) prend la forme

$$\frac{u_m^{p+1} - u_m^p}{\tau} + \frac{1}{h}\left[\frac{(u_m^p)^2}{2} - \frac{(u_{m-1}^p)^2}{2}\right] = 0,$$

$$u_m^0 = \frac{1}{h}\int_{x_{m-1/2}}^{x_{m+1/2}} \psi(x)\,dx$$

(*)La fonction $u = u(x, t)$ est définie presque partout et la fonction $u^{(h)} = u^{(h)}(x, t)$ seulement sur un réseau de droites. On vient à bout de cette discordance formelle en supposant qu'avec la diminution de h chaque nouveau réseau est une partie de l'ancien et en parlant de la convergence aux points du réseau construit pour tout h fixe admissible.

ou

$$\frac{u_m^{p+1}-u_m^p}{\tau}+\left(\frac{u_{m-1}^p+u_m^p}{2}\right)\frac{u_m^p-u_{m-1}^p}{h}=0.$$

On voit aisément que pour

$$r=\frac{\tau}{h}\leqslant\frac{1}{\max\limits_m|u_m^p|}$$

joue le principe du maximum

$$\max_m|u_m^{p+1}|\leqslant\max_m|u_m^p|\leqslant\ldots\leqslant\max_m|u_m^0|\leqslant\max_x|\psi(x)|.$$

Il en résulte que pour $\tau=\dfrac{1}{\max|\psi(x)|}h$ on peut espérer que le schéma obtenu est stable pour un certain choix raisonnable des normes. Nous nous abstiendrons d'indiquer celles-ci : les calculs expérimentaux confirment que, le réseau se resserrant, la solution $u^{(h)}$ du problème (7) avec les données initiales $\psi(x)$ régulières par morceaux et monotones par morceaux converge vers une fonction $u(x, t)$ présentant des discontinuités en nombre fini et que la convergence est uniforme à l'extérieur de tout voisinage des discontinuités.

Le schéma (7) où $U_{m+1/2}^{p+1/2}$ se calcule moyennant la suppression de la discontinuité n'est évidemment pas l'unique schéma conservatif pour le problème (1). En voici un autre, très simple, dont le principe est la prédiction-correction, technique exposée au n° 3 du §22. Pour ne pas alourdir nos raisonnements, bornons-nous au cas $\psi(x)>0$.

Cherchons d'abord les quantités auxiliaires $\bar{u}$ par le schéma implicite non conservatif

$$\frac{\bar{u}_m^{p+1/2}-u_m^p}{\tau/2}+u_m^p\,\frac{\bar{u}_m^{p+1/2}-\bar{u}_{m-1}^{p+1/2}}{h}=0.$$

En substituant à la valeur du coefficient de u_x de l'équation $u_t+uu_x=0$ la quantité u_m^p et non pas $\bar{u}_m^{p+1/2}$ nous avons voulu obtenir un schéma linéaire en des quantités à calculer.

Posons ensuite

$$U_{m+1/2}^{p+1/2}=\frac{1}{2}\left(\bar{u}_m^{p+1/2}+\bar{u}_{m+1}^{p+1/2}\right)\tag{9}$$

et utilisons le schéma (7), (9).

Le schéma conservatif obtenu est d'ordre d'approximation deux pour la solution régulière.

Une étude heuristique avec recours au critère spectral de von Neumann, à la linéarisation et à la fixation des coefficients révèle la stabilité pour $r=\tau/h$ arbitraire.

Par linéarisation et fixation des coefficients, nous aboutissons à un

schéma de la forme

$$\left.\begin{aligned} \frac{\bar{u}_m^{p+1/2}-u_m^p}{\tau/2}+a\,\frac{\bar{u}_m^{p+1/2}-\bar{u}_{m-1}^{p+1/2}}{h} &= 0,\\ \frac{u_m^{p+1}-u_m^p}{\tau}+\frac{a}{h}\left[\frac{\bar{u}_{m+1}^{p+1/2}+\bar{u}_m^{p+1/2}}{2}-\frac{\bar{u}_m^{p+1/2}+\bar{u}_{m-1}^{p+1/2}}{2}\right] &= 0. \end{aligned}\right\}$$

En résolvant avec les données initiales

$$u_m^p = e^{i\alpha m}$$

nous obtenons

$$\bar{u}_m^{p+1/2} = \mu e^{i\alpha m},$$

où

$$\mu = \frac{1}{1+a\dfrac{r}{2}-a\dfrac{r}{2}\,e^{-i\alpha}}.$$

Ensuite,

$$u_m^{p+1} = \lambda e^{i\alpha m},$$

où

$$\frac{\lambda-1}{\tau}+\frac{a\mu}{h}\left(\frac{e^{i\alpha}-e^{-i\alpha}}{2}\right) = 0,$$

$$\lambda(\alpha) = \frac{2+ar-are^{+i\alpha}}{2+ar-are^{-i\alpha}}, \quad |\lambda(\alpha)| = 1.$$

QUATRIÈME PARTIE

PROBLÈMES À DEUX VARIABLES D'ESPACE

CHAPITRE 10

NOTION DE SCHÉMAS DE DÉCOMPOSITION AUX DIFFÉRENCES

Les schémas de décomposition aux différences sont un outil important du calcul de solutions des problèmes instationnaires pluridimensionnels de physique mathématique.

§31. Construction de schémas de décomposition

Au niveau descriptif l'idée à la base de la construction de schémas de décomposition est la suivante.

Soit un problème différentiel de la forme

$$\left.\begin{aligned} \frac{\partial u}{\partial t} &= Au, \quad 0 < t < T, \\ u|_{t=0} &\text{ donné,} \end{aligned}\right\} \tag{1}$$

avec A un opérateur sur les variables d'espace, par exemple :

$$Au \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Les valeurs $u(x, y, t_{p+1})$ s'expriment en fonction des valeurs connues $u(x, y, t_p)$, $t_p = p\tau$, par la formule

$$u(x, y, t_p+\tau) = u(x, y, t_p) + \tau \frac{\partial u}{\partial t} + O(\tau^2) =$$
$$= u(x, y, t_p) + \tau Au(x, y, t_p) + O(\tau^2) = (E+\tau A)\, u(x, y, t_p) + O(\tau^2),$$

où E est l'opérateur unité, $Ev = v$.

Admettons que le second membre de l'équation (1) est de la forme

$$Au \equiv A_1 u + A_2 u.$$

Décomposons alors l'équation (1)

$$\frac{\partial u}{\partial t} = A_1 u + A_2 u$$

en deux équations suivantes :

$$\left.\begin{aligned} \frac{\partial v}{\partial t} &= A_1 v, \quad t_p \leqslant t \leqslant t_{p+1}, \\ v(x, y, t_p) &= u(x, y, t_p), \end{aligned}\right\} \tag{2}$$

$$\left.\begin{aligned} \frac{\partial w}{\partial t} &= A_2 w, \quad t_p \leqslant t \leqslant t_{p+1}, \\ w(x, y, t_p) &= v(x, y, t_{p+1}). \end{aligned}\right\} \tag{3}$$

Notons que

$$w(x, y, t_{p+1}) = u(x, y, t_{p+1}) + O(\tau^2). \tag{4}$$

En effet,

$$v(x, y, t_{p+1}) = (E+\tau A_1)\, v(x, y, t_p) + O(\tau^2) = (E+\tau A_1)\, u(x, y, t_p) + O(\tau^2).$$

Compte tenu de la dernière égalité,

$$\begin{aligned} w(x, y, t_{p+1}) &= (E+\tau A_2)\, w(x, y, t_p) + O(\tau^2) = (E+\tau A_2)\, v(x, y, t_{p+1}) + \\ &+ O(\tau^2) = (E+\tau A_2)(E+\tau A_1)\, u(x, y, t_p) + O(\tau^2) = \\ &= [E+\tau(A_1+A_2)]\, u(x, y, t_p) + \tau^2 A_1 A_2 u(x, y, t_p) + \\ &+ O(\tau^2) = (E+\tau A)\, u(x, y, t_p) + O(\tau^2) = u(x, y, t_{p+1}) + O(\tau^2). \end{aligned}$$

En vertu de l'égalité (4) nous avons le droit de résoudre successivement dans chaque intervalle temporel $t_p \leqslant t \leqslant t_{p+1}$ les problèmes (2) et (3) au lieu de (1).

La résolution réelle de (2) et (3) exige qu'on les approche formellement par des équations aux différences. Il apparaît alors un schéma de décomposition aux différences $L_h u^{(h)} = f^{(h)}$ permettant de calculer u^{p+1} en deux étapes d'après u^p connu (on calcule en première étape v^{p+1} en fonction de $v^p = u^p$ donné, et en seconde étape on recherche $u^{p+1} = w^{p+1}$ à partir de $w^p = v^{p+1}$ obtenu à l'étape précédente).

Il s'agit là de raisonnements heuristiques. Après avoir construit le schéma de décomposition

$$L_h u^{(h)} = f^{(h)} \tag{5}$$

servant à résoudre le problème (1), il faut vérifier d'une façon ou d'une autre son approximation et sa stabilité.

Dans le cas du problème de Cauchy pour l'équation de la chaleur à deux dimensions

$$\left.\begin{aligned} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad 0 < t < T, \quad -\infty < x, \quad y < \infty, \\ u(x, y, 0) = \psi(x, y) \end{aligned}\right\} \tag{6}$$

on peut prendre pour système (2), (3)

$$\left.\begin{aligned} \frac{\partial v}{\partial t} &= \frac{\partial^2 v}{\partial x^2}, \quad v(x, y, t_p) = u(x, y, t_p), \\ \frac{\partial w}{\partial t} &= \frac{\partial^2 w}{\partial y^2}, \quad w(x, y, t_p) = v(x, y, t_{p+1}). \end{aligned}\right\} \tag{7}$$

Cette décomposition de l'équation bidimensionnelle du problème (6) en deux équations unidimensionnelles (7) peut s'interpréter comme une substitution approchée de deux processus au processus de propagation de la chaleur dans le plan Oxy durant $t_p \leqslant t \leqslant t_{p+1}$. On introduit par la pensée dans le premier processus partiel décrit par la première équation (7) des parois calorifuges interdisant la propagation de la chaleur dans la direction de l'axe Oy. Au bout d'une période τ ces parois sont remplacées par d'autres empêchant la chaleur de se propager dans la direction de l'axe Ox. La propagation de la chaleur est alors décrite, durant un autre intervalle τ, par la seconde équation.

Choisissons le réseau $(x_m, y_n, t_p) = (mh, nh, p\tau)$.

Compte tenu de (7), la construction du schéma de décomposition aux différences se fait de différentes manières. Indiquons-en deux :

$$\left.\begin{aligned} \frac{\tilde{u}_{mn} - u^p_{mn}}{\tau} &= \Lambda_{xx} u^p_{mn}, \\ \frac{u^{p+1}_{mn} - \tilde{u}_{mn}}{\tau} &= \Lambda_{yy} \tilde{u}_{mn}, \\ u^0_{mn} &= \psi(x_m, y_n) \end{aligned}\right\} \tag{8}$$

et

$$\left.\begin{aligned} \frac{\tilde{u}_{mn} - u^p_{mn}}{\tau} &= \Lambda_{xx} \tilde{u}_{mn}, \\ \frac{u^{p+1}_{mn} - \tilde{u}_{mn}}{\tau} &= \Lambda_{yy} u^{p+1}_{mn}, \\ u^0_{mn} &= \psi(x_m, y_n). \end{aligned}\right\} \tag{9}$$

Posons pour les deux schémas

$$\tilde{u}_{mn} \equiv v^{p+1}_{mn} \equiv w^p_{mn}, \quad u^{p+1}_{mn} \equiv w^{p+1}_{mn}.$$

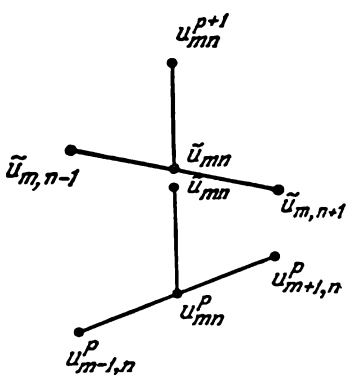


Fig. 40

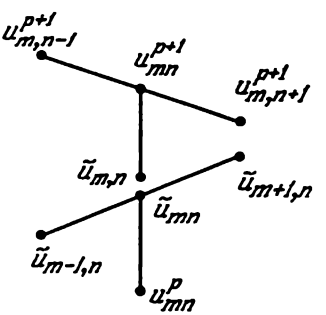


Fig. 41

Rappelons les désignations suivantes :

$$\Lambda_{xx}u_{mn} = \frac{u_{m+1,n} - 2u_{mn} + u_{m-1,n}}{h^2},$$

$$\Lambda_{yy}u_{mn} = \frac{u_{m,n+1} - 2u_{mn} + u_{m,n-1}}{h^2}.$$

Le schéma (8) est représenté à la figure 40, le schéma (9) à la figure 41.

La décomposition du problème (6) n'est pas unique elle non plus. Celui-ci peut s'écrire comme suit :

$$\left.\begin{aligned} \frac{\partial u}{\partial t} &= \frac{1}{2}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + \frac{1}{2}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right), \\ u(x, y, 0) &= \psi(x, y) \end{aligned}\right\} \tag{6'}$$

et on peut lui faire correspondre sur le segment $t_p \leqslant t \leqslant t_{p+1}$ deux systèmes suivants :

$$\left.\begin{aligned} \frac{\partial v}{\partial t} &= \frac{1}{2}\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right), \quad t_p \leqslant t \leqslant t_{p+1}, \\ v(x, y, t_p) &= u(x, y, t_p) \end{aligned}\right\} \tag{10}$$

et

$$\left.\begin{aligned} \frac{\partial w}{\partial t} &= \frac{1}{2}\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2}\right), \quad t_p \leqslant t \leqslant t_{p+1}, \\ w(x, y, t_p) &= v(x, y, t_{p+1}). \end{aligned}\right\} \tag{11}$$

Ce n'est pas une décomposition suivant des considérations physiques comme pour le schéma (7). Le schéma aux différences est choisi comme suit (fig. 42) :

$$\left.\begin{aligned} \frac{\tilde{u}_{mn} - u^p_{mn}}{\tau} &= \frac{1}{2}(\Lambda_{xx}u^p_{mn} + \Lambda_{yy}\tilde{u}_{mn}), \\ \frac{u^{p+1}_{mn} - \tilde{u}_{mn}}{\tau} &= \frac{1}{2}(\Lambda_{xx}u^{p+1}_{mn} + \Lambda_{yy}\tilde{u}_{mn}), \\ u^0_{mn} &= \psi(x_m, y_n). \end{aligned}\right\} \tag{12}$$

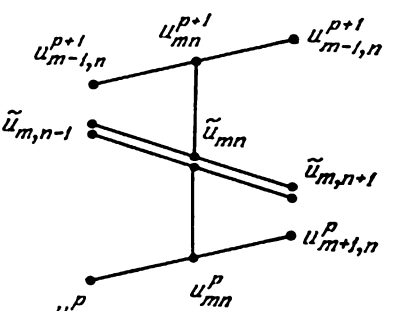


Fig. 42

Dans le calcul de u^{p+1} par le *schéma de directions alternées* (12) on résout d'abord pour chaque m fixe l'équation implicite par rapport à $\tilde{u}_{mn}$ contenant m comme paramètre. On calcule ensuite u^{p+1}_{mn} en résolvant la seconde équation (12), qui est implicite en u^{p+1}_{mn} et où n figure en tant que paramètre.

Le schéma (8) prend la forme (5) à condition de poser

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u^{p+1}_{mn} - \tilde{u}_{mn}}{\tau} - \Lambda_{yy}\tilde{u}_{mn}, \\ u^0_{mn}, \end{cases}$$

avec $\tilde{u}_{mn} = u^p_{mn} + \tau\Lambda_{xx}u^p_{mn}$ défini à partir de la première équation (8). On a alors

$$f^{(h)} \equiv \begin{cases} 0, & m, n = 0, \pm 1, \ldots; \quad p = 0, 1, \ldots, [T/\tau]-1, \\ \psi(x_m, y_n), & m, n = 0, \pm 1, \ldots \end{cases}$$

Nous laissons au lecteur le soin de mettre les schémas (9) et (12) sous la forme (5).

Le lecteur peut vérifier que le critère spectral de stabilité de von Neumann qui exige que les solutions de la forme $u^p_{mn} = \lambda^p e^{i(\alpha m+\beta n)}$ soient bornées, est satisfait pour le schéma (8) avec $r = \tau/h^2 \leqslant {}^1/_2$ et pour les schémas (9) et (12) avec r quelconque.

Nous nous abstiendrons d'étudier les conditions de stabilité et de démontrer l'approximation des schémas (8), (9) et (12).

Exercices

1. Etablir pour quels $r = \tau/h^2$ a lieu le critère spectral de von Neumann pour les schémas (8), (9) et (12) étudiés ci-dessus.
2. Vérifier que le schéma (8) approche le problème (6) pour une solution bornée suffisamment régulière $u(x, y, t)$.
3. Même question que dans l'exercice 2 pour les schémas (9) et (12).

§32. Schémas aux différences économiques

Discutons les exemples de schémas de décomposition aux différences pour le problème de propagation de la chaleur

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \qquad 0 \leqslant x, \quad y \leqslant 1, \quad 0 \leqslant t \leqslant T, \\ &u(x, y, 0) = \psi(x, y), \quad 0 \leqslant x, \quad y \leqslant 1, \\ &u(x, y, t)|_\Gamma = 0 \end{aligned}\right\} \tag{1}$$

dans le domaine rectangulaire $0 \leqslant x$, $y \leqslant 1$ de frontière Γ. Utilisons le réseau habituel $(x_m, y_n, t_p) = (mh, nh, p\tau)$, $m, n = 1, 2, \ldots, N$ et $h = 1/N$.

Le schéma de décomposition aux différences que nous allons donner présente sous certains rapports des avantages essentiels sur le plus simple schéma explicite

$$\left.\begin{aligned} \frac{u^{p+1}_{mn} - u^p_{mn}}{\tau} &= \Lambda_{xx}u^p_{mn} + \Lambda_{yy}u^p_{mn}, \\ u^0_{mn} &= \psi(x_m, y_n), \\ u^p_{mn}|_\Gamma &= 0 \end{aligned}\right\} \tag{2}$$

et celui implicite

$$\left.\begin{aligned} \frac{u^{p+1}_{mn} - u^p_{mn}}{\tau} &= \Lambda_{xx}u^{p+1}_{mn} + \Lambda_{yy}u^{p+1}_{mn}, \\ u^0_{mn} &= \psi(x_m, y_n), \\ u^p_{mn}|_\Gamma &= 0. \end{aligned}\right\} \tag{3}$$

Le calcul selon le schéma aux différences explicite (2) est fort simple. Le passage de u^p connu à $u^{p+1} = \{u_{mn}^{p+1}\}$ cherché exige des opérations arithmétiques en nombre proportionnel au nombre $(N-1)^2$ de valeurs inconnues de $\{u_{mn}^{p+1}\}$. En ce sens le schéma explicite ne se prête à aucune amélioration. Des schémas aux différences tels que le nombre d'opérations arithmétiques nécessaires pour passer de u^p à $u^{p+1} = \{u_{mn}^{p+1}\}$ soit proportionnel à celui de valeurs inconnues sont dits *économiques*. Tout en étant économique, le schéma explicite n'est stable que pour la contrainte sévère $\tau \leqslant h^2/4$ sur le pas τ du réseau. La stabilité absolue (§27, n° 3) du schéma aux différences implicite très simple (3) ne signifie nullement son caractère économique. La recherche des inconnues $\{u_{mn}^{p+1}\}$ se fait en résolvant un système d'équations linéaires compliqué (qui ne se décompose pas). On sait du cours d'algèbre que cela demande des opérations arithmétiques en nombre proportionnel non pas à la puissance première du nombre d'inconnues (cas des schémas économiques), mais au cube de ce nombre si l'on se sert d'une méthode d'élimination des inconnues.

On recherche actuellement des procédés plus économiques de résolution exacte des systèmes linéaires généraux. Un algorithme dû à Strassen coûte un nombre d'opérations arithmétiques qui est proportionnel non pas à la puissance troisième du nombre d'inconnues mais à la $\lg_2 7$-ième.

Le schéma de décomposition aux différences que nous nous proposons de construire est économique et inconditionnellement stable, i.e. il réunit les avantages du schéma explicite (2) à ceux du schéma implicite (3).

En ce qui concerne la solution $u(x, y, t)$ du problème (1), elle sera supposée posséder jusqu'à sur la frontière Γ les dérivées continues d'ordre quelconque dont nous aurons besoin. Notons que sur Γ toutes les dérivées d'ordre pair (jusqu'à l'ordre pour lequel elles existent et sont continues) par rapport aux variables d'espace s'annulent :

$$u_{xx}|_\Gamma = u_{xxxx}|_\Gamma = u_{xxyy}|_\Gamma = 0. \tag{4}$$

Ainsi, sur le côté $x = 0$ du carré $0 \leqslant x, y \leqslant 1$ ce sont $\partial u/\partial t$ et $\partial^2 u/\partial y^2$ qui jouissent de cette propriété. En vertu de l'équation $u_t = u_{xx} + u_{yy}$ on a donc également $u_{xx} = 0$. Dérivons deux fois par rapport à y, il vient

$$\frac{\partial u_{yy}}{\partial t} = u_{xxyy} + u_{yyyy}.$$

Mais sur le côté $x = 0$

$$u_{yy} = 0, \quad u_{yyyy} = 0, \quad \frac{\partial u_{yy}}{\partial t} = 0$$

et donc $u_{xxyy} = 0$ en vertu de l'équation différentielle.

Construisons à présent un schéma de décomposition aux différences pour le problème (1).

Faisons correspondre à celui-ci sur le segment $t_p \leqslant t \leqslant t_{p+1}$ deux

problèmes :

$$\left.\begin{aligned} &\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2}, \\ &v|_\Gamma = 0, \quad v(x, y, t_p) = u(x, y, t_p), \end{aligned}\right\} \tag{5}$$

$$\left.\begin{aligned} &\frac{\partial w}{\partial t} = \frac{\partial^2 w}{\partial y^2}, \\ &w|_\Gamma = 0, \quad w(x, y, t_p) = v(x, y, t_{p+1}). \end{aligned}\right\} \tag{6}$$

La fonction

$$u^{(h)} = \{u_{mn}^p\}, \quad u_{mn}^0 = \psi(x_m, y_n), \quad u_{mn}^p|_\Gamma = 0$$

sera définie successivement pour $p = 1, 2, \ldots, [T/\tau]$ par les équations

$$\left.\begin{aligned} &\frac{\tilde{u}_{mn} - u_{mn}^p}{\tau} = \Lambda_{xx}\tilde{u}_{mn}, \quad m, n = 1, 2, \ldots, N-1, \\ &\tilde{u}_{mn}|_\Gamma = 0; \end{aligned}\right\} \tag{7}$$

$$\left.\begin{aligned} &\frac{u_{mn}^{p+1} - \tilde{u}_{mn}}{\tau} = \Lambda_{yy}u_{mn}^{p+1}, \quad m, n = 1, 2, \ldots, N-1, \\ &u_{mn}^{p+1}|_\Gamma = 0. \end{aligned}\right\} \tag{8}$$

Le problème (7) est analogue à (5) et le problème (8) à (6). Cela étant,

$$v_{mn}^p = u_{mn}^p, \quad w_{mn}^p = v_{mn}^{p+1} = \tilde{u}_{mn}, \quad w_{mn}^{p+1} = u_{mn}^{p+1}.$$

Conformément au schéma de décomposition (7), (8) on calcule d'abord la fonction auxiliaire $\tilde{u}_{mn}$ d'après les valeurs connues $u^p = \{u_{mn}^p\}$, puis $u^{p+1} = \{u_{mn}^{p+1}\}$ par l'équation (8).

Notons que le problème aux différences (7) servant à obtenir $\tilde{u}_{mn}$ pour chaque n fixe, $n = 1, \ldots, N-1$, coïncide exactement avec le schéma aux différences implicite

$$\frac{v_{mn}^{p+1} - v_{mn}^p}{\tau} = \Lambda_{xx}v_{mn}^{p+1}$$

pour l'équation de la chaleur à une dimension (5) sur le segment $0 \leqslant x \leqslant 1$ où y figure seulement comme paramètre.

Le problème aux différences (7) se résout pour tout n fixe par balayage dans la direction de l'axe Ox.

De même, le problème (8) est résolu pour tout m fixe par balayage dans la direction de l'axe Oy. Soulignons qu'en vertu des propriétés de l'algorithme de balayage, le nombre total d'opérations arithmétiques pour calculer $u^{p+1} = \{u_{mn}^{p+1}\}$ est proportionnel au nombre de valeurs inconnues, ce qui fait que le schéma aux différences (7), (8) est économique.

Si l'on veut formuler rigoureusement les notions d'approximation et de stabilité, il faut écrire le schéma aux différences (7), (8) sous la forme

$$L_h u^{(h)} = f^{(h)}, \tag{9}$$

forme que nous utilisons partout dans le présent ouvrage. Posons à cette

fin

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{mn}^{p+1} - \tilde{u}_{mn}}{\tau} - \Lambda_{yy} u_{mn}^{p+1}, & m, n = 1, 2, \ldots, N-1, \\ u_{mn}^{p+1}|_{\Gamma}, \\ u_{mn}^{0}, \end{cases} \tag{10}$$

où $\tilde{u}_{mn}$ est solution du problème auxiliaire

$$\left.\begin{aligned} \frac{\tilde{u}_{mn} - u_{mn}^{p}}{\tau} &= \Lambda_{xx} \tilde{u}_{mn}, \quad m, n = 1, 2, \ldots, N-1, \\ \tilde{u}_{mn}|_{\Gamma} &= 0. \end{aligned}\right\} \tag{11}$$

On prend alors pour $f^{(h)}$

$$f^{(h)} \equiv \begin{cases} 0, & m, n = 1, 2, \ldots, N-1, \\ 0, & (x_m, y_n) \in \Gamma, \\ \psi(x_m, y_n) \end{cases} \tag{12}$$

et pour norme dans U_h

$$\|u^{(h)}\|_{U_h} = \max_{m, n, p} |u_{mn}^{p}|.$$

L'espace F_h sera supposé contenir des éléments $g^{(h)}$ de la forme

$$g^{(h)} = \begin{cases} \varphi_{mn}^{p}, \\ 0, \\ \psi_{mn} \end{cases}$$

et la norme dans F_h se définira par l'égalité

$$\|g^{(h)}\|_{F_h} = \max_{m, n, p} |\varphi_{mn}^{p}| + \max_{m, n} |\psi_{mn}|.$$

Démontrons en premier lieu la stabilité inconditionnelle du schéma aux différences (9) donné par les équations (10), (12). Le schéma (9) étant linéaire, la stabilité se démontre à condition de montrer que le problème $L_h z^{(h)} = g^{(h)}$ possède une solution pour $g^{(h)} = \begin{Bmatrix} \varphi_{mn}^{p} \\ 0 \\ \psi_{mn} \end{Bmatrix} \in F_h$ arbitraire, de plus

$$\|z^{(h)}\|_{U_h} \leqslant c \|g^{(h)}\|_{F_h},$$

avec c indépendant de h.

Ecrivons le problème $L_h z^{(h)} = g^{(h)}$ sous forme développée

$$\left.\begin{aligned} \frac{z_{mn}^{p+1} - \tilde{z}_{mn}}{\tau} - \Lambda_{yy} z_{mn}^{p+1} &= \varphi_{mn}^{p+1}, \\ m, n &= 1, 2, \ldots, N-1, \\ z_{0n}^{p+1} = z_{Nn}^{p+1} &= 0, \end{aligned}\right\} \tag{13}$$

où $\tilde{z}_{mn}$ est solution du problème auxiliaire

$$\left.\begin{aligned} &\frac{\tilde{z}_{mn}-z^p_{mn}}{\tau} = \Lambda_{xx}\tilde{z}_{mn}, \quad m, n = 1, 2, \ldots, N-1, \\ &\tilde{z}_{m,0} = \tilde{z}_{m,N} = 0, \end{aligned}\right\} \tag{14}$$

et

$$z^0_{mn} = \psi_{mn}.$$

En vertu du principe du maximum démontré (§28) pour un schéma aux différences implicite unidimensionnel qui approche l'équation de la chaleur à une dimension, il résulte des équations (13) que

$$\max_{m,n} |z^{p+1}_{mn}| \leqslant \max_{m,n} |\tilde{z}_{mn}| + \tau \max_{m,n,p} |\varphi^p_{mn}|.$$

Or, (14) entraîne

$$\max_{m,n} |\tilde{z}_{mn}| \leqslant \max_{m,n} |z^p_{mn}|.$$

Donc

$$\max_{m,n} |z^{p+1}_{mn}| \leqslant \max_{m,n} |z^p_{mn}| + \tau \max_{m,n,p} |\varphi^p_{mn}|.$$

D'où

$$\max_{m,n} |z^{p+1}_{mn}| \leqslant \max_{m,n} |z^{p-1}_{mn}| + 2\tau \max |\varphi^p_{mn}| \leqslant$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\leqslant \max_{m,n} |\psi_{mn}| + T \max_{m,n,p} |\varphi^p_{mn}| \leqslant (1+T)\Big(\max_{m,n} |\psi_{mn}| +$$

$$+ \max_{m,n,p} |\varphi^p_{mn}|\Big) = (1+T)\,||g^{(h)}||_{F_h}.$$

L'inégalité écrite

$$\max_{m,n} |z^{p+1}_{mn}| \leqslant (1+T)\,||g^{(h)}||_{F_h}$$

est juste quel que soit p, aussi

$$||z^{(h)}||_{U_h} \leqslant (1+T)\,||g^{(h)}||_{F_h}$$

pour un rapport arbitraire des pas τ et h. Cela signifie justement la stabilité inconditionnelle du schéma de décomposition considéré.

Vérifions l'approximation et supposons comme toujours que le problème (1) admet une solution suffisamment régulière $u(x, y, t)$. Calculons le résidu $\Delta f^{(h)}$, $L_h[u]_h = f^{(h)} + \Delta f^{(h)}$, résultant de la substitution de $[u]_h$ dans le premier membre de (9) et montrons que $||\Delta f^{(h)}||_{F_h} = O(\tau + h^2)$.

Conformément à la définition de l'opérateur L_h, nous aurons

$$L_h[u]_h = \begin{cases} \dfrac{u(x_m, y_n, t_{p+1}) - \tilde{u}_{mn}}{\tau} - \Lambda_{yy} u(x_m, y_n, t_{p+1}), \\ \qquad m, n = 1, \ldots, N-1, \\ 0 \text{ aux points de } \Gamma, \\ u(x_m, y_n, 0), \quad m, n = 1, \ldots, N-1, \end{cases} \tag{15}$$

$\tilde{u}_{mn}$ étant solution du problème auxiliaire

$$\left.\begin{aligned} &\frac{\tilde{u}_{mn}-u(x_m, y_n, t_p)}{\tau}-\Lambda_{xx}\tilde{u}_{mn} = 0, \quad m, n = 1, 2, \ldots, N-1, \\ &\tilde{u}_{mn}|_\Gamma = 0. \end{aligned}\right\} \tag{16}$$

Nous montrerons plus loin que $\tilde{u}_{mn}$ est de la forme

$$\left.\begin{aligned} &\tilde{u}_{mn} = u(x_m, y_n, t_p)+\tau\Lambda_{xx}u(x_m, y_n, t_p)+O(\tau^2), \\ &\qquad m, n = 1, 2, \ldots, N-1, \\ &\tilde{u}_{mn}|_\Gamma = u(x_m, y_n, t_p)|_\Gamma = 0. \end{aligned}\right\} \tag{17}$$

Le report de cette expression de $\tilde{u}_{mn}$ dans (15) donne

$$\frac{u(x_m, y_n, t_{p+1})-\tilde{u}_{mn}}{\tau}-\Lambda_{yy}u(x_m, y_n, t_{p+1}) =$$

$$= \frac{u(x_m, y_n, t_{p+1})-[u(x_m, y_n, t_p)+\tau\Lambda_{xx}u(x_m, y_n, t_p)+O(\tau^2)]}{\tau}-$$

$$-\Lambda_{yy}u(x_m, y_n, t_{p+1}) = \frac{u(x_m, y_n, t_{p+1})-u(x_m, y_n, t_p)}{\tau}-$$

$$-\Lambda_{xx}u(x_m, y_n, t_p)-\Lambda_{yy}u(x_m, y_n, t_{p+1})+O(\tau) =$$

$$= \left(\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}-\frac{\partial^2 u}{\partial y^2}\right)_{x_m, y_n, t_p}+O(\tau+h^2) = O(\tau+h^2).$$

Ainsi,

$$L_h[u_h] \equiv \begin{Bmatrix} 0+O(\tau+h^2) \\ 0+0 \\ \psi(x_m, y_n)+0 \end{Bmatrix} = f^{(h)}+\Delta f^{(h)},$$

où

$$\Delta f^{(h)} = \begin{cases} O(\tau+h^2) & \text{aux points } (x_m, y_n, t_p), \\ 0, & p = 0, 1, \ldots, [T/\tau]-1, \\ 0, & m, n = 1, 2, \ldots, N-1. \end{cases}$$

D'où

$$\|\Delta f^{(h)}\|_{F_h} = O(\tau+h^2).$$

Il nous reste à démontrer la représentation approchée (17) de la solution $\tilde{u}_{mn}$ du problème (16).

Enonçons d'abord les considérations heuristiques qui nous ont suggéré la représentation (17). Il est clair que pour des τ petits on a l'égalité approchée

$$\tilde{u}_{mn} \approx u(x_m, y_n, t_p).$$

Cela nous permet de remplacer dans (16) $\Lambda_{xx}\tilde{u}_{mn}$ par $\Lambda_{xx}u(x_m, y_n, t_p)$ et d'obtenir donc

$$\frac{\tilde{u}-u}{\tau}-\Lambda_{xx}u = 0.$$

L'égalité $\tilde{u} = u+\tau\Lambda_{xx}u$ qui en découle ne diffère de (17) que par le reste $O(\tau^2)$. Passons à la démonstration de (17).

Définissons $\Lambda_{xx}u(x_m, y_n, t_p)$ aux points de Γ en posant $\Lambda_{xx}u|_\Gamma = 0$. Substituons

$$w_{mn} \equiv u(x_m, y_n, t_p)+\tau\Lambda_{xx}u(x_m, y_n, t_p)$$

à $\tilde{u}_{mn}$ dans l'équation (16). Il vient

$$\frac{w_{mn}-u(x_m, y_n, t_p)}{\tau}-\Lambda_{xx}w_{mn} \equiv$$

$$\equiv \frac{\{u(x_m, y_n, t_p)+\tau\Lambda_{xx}u(x_m, y_n, t_p)\}-u(x_m, y_n, t_p)}{\tau}-\Lambda_{xx}u(x_m, y_n, t_p)-$$

$$-\tau\Lambda_{xx}\Lambda_{xx}u(x_m, y_n, t_p) = -\tau\Lambda_{xx}\Lambda_{xx}u(x_m, y_n, t_p).$$

Sous l'hypothèse de $\partial^4u/\partial x^4$ continu et borné et vu que $\left.\frac{\partial^2 u}{\partial x^2}\right|_\Gamma = 0$, on constate sans peine que $\Lambda_{xx}\Lambda_{xx}u(x_m, y_n, t_p)$ est borné lui aussi. Donc

$$\frac{w_{mn}-u(x_m, y_n, t_p)}{\tau}-\Lambda_{xx}w_{mn} = O(\tau),$$

$$w_{mn}|_\Gamma = 0.$$

La soustraction terme à terme des équations (16) donne pour la différence $z_{mn} = w_{mn}-\tilde{u}_{mn}$:

$$\left.\begin{aligned} z_{mn}-\tau\Lambda_{xx}z_{mn} &= O(\tau^2), \\ z_{mn}|_\Gamma &= 0, \end{aligned}\right\} \tag{18}$$

ou sous forme développée

$$\left.\begin{aligned} rz_{m-1,n}-2\left(r+\frac{1}{2}\right)z_{mn}+rz_{m+1,n} &= O(\tau^2), \\ m, n &= 1, \ldots, N-1, \\ z_{0n} = z_{Mn} = 0, \quad r &= \tau/h^2. \end{aligned}\right\} \tag{18'}$$

Ce problème pour $\{z_{mn}\}$ prend la forme

$$a_mu_{m-1}+b_mu_m+c_mu_{m+1} = g_m, \quad m = 1, \ldots, N-1,$$

$$u_0 = u_N = 0,$$

$$a_m > 0, \quad c_m > 0, \quad |b_m| \geqslant a_m+c_m+\delta, \quad \delta = 1.$$

Il a été démontré au §4 que dans ce cas

$$\max|u_m| \leqslant c \max|g_m|,$$

c dépendant de δ seul. D'où $z_{mn} = O(\tau^2)$, i.e.

$$\tilde{u}_{mn} = w_{mn}-z_{mn} = w_{mn}+O(\tau^2) =$$

$$= u(x_m, y_n, t_p)+\tau\Lambda_{xx}u(x_m, y_n, t_p)+O(\tau^2),$$

ce qui coïncide avec la représentation (17), objet de notre démonstration.

Exercices

1. Pour le problème aux limites différentiel (1) relatif à la propagation de la chaleur dans le domaine carré $0 \leqslant x, y \leqslant 1$ proposer et discuter un schéma de décomposition aux différences analogue au schéma explicite (8) pour le problème de Cauchy (§31).

2. Proposer pour le problème aux limites différentiel (1) un schéma aux différences analogue au schéma de directions alternées (12) du §31. Démontrer qu'il y a lieu une approximation d'ordre $O(\tau+h^2)$.

3. Proposer un schéma aux différences pour la résolution du problème

$$\left.\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad 0 \leqslant t \leqslant T, \quad (x, y) \in D \\ &u(x, y, t)|_\Gamma = \psi(x, y, t), \quad u(x, y, 0) = \varphi(x, y) \end{aligned}\right\}$$

dans un domaine curviligne D par analogie avec le schéma de décomposition aux différences considéré pour le problème (1) dans le présent paragraphe.

§33. Décomposition suivant les facteurs physiques

L'idée de décomposition ne sert pas qu'à obtenir des schémas absolument stables économiques. On décompose parfois un problème compliqué en une succession de problèmes plus simples afin de séparer dans le temps dans chaque petit intervalle $t_p < t < t_{p+1}$ l'action de divers facteurs influant sur le processus donné. Il est en effet plus facile de construire, pour ces problèmes relativement simples, des schémas adéquats dont l'ensemble constitue un schéma de décomposition aux différences raisonnable relatif à tout le problème.

Indiquons à ce propos la méthode de grosses particules de Bélotserkovski et Davydov (Zh. Vychisl. Mat. i Mat. Fiz. **11**, n° 1, 1971) destinée à calculer les écoulements gazeux en présence d'une déformation importante de la substance et de grandes oscillations de densité. Cette méthode, de même que la PIC (the particule-in-cell method) de Harlow, peut s'interpréter, selon Yanenko, comme un schéma de décomposition aux différences pour des équations de la mécanique des fluides. La substance toute entière est partagée en cellules par un réseau de droites fixes (il s'agit d'un problème plan). La substance venue dans une cellule à l'instant t_p est justement une grosse particule. On la dote d'une impulsion et de l'énergie totale. On construit ensuite un schéma aux différences qui sert à reproduire la variation des vitesses, des impulsions et de l'énergie totale des grosses particules sous l'influence de la seule pression, abstraction faite des termes du système d'équations de la mécanique des fluides qui gouvernent le transport de substance, d'impulsion et d'énergie. Telle est la première étape du schéma de décomposition. En seconde étape on recalcule les quantités auxiliaires obtenues à l'étape précédente à l'aide d'un schéma aux différences tenant compte des autres termes des équations, i.e. traduisant le passage de la substance d'une cellule dans les cellules voisines et le transport correspondant d'impulsion et d'énergie. On obtient ainsi de grosses particules et l'impulsion et l'énergie correspondantes à l'instant $t_{p+1} = t_p+\tau$.

CHAPITRE 11

PROBLÈMES ELLIPTIQUES

§34. Un schéma aux différences très simple pour le problème de Dirichlet

Nous vérifierons dans ce paragraphe que le schéma aux différences très simple (13) du §24 approche le problème de Dirichlet (12) du même paragraphe à l'ordre deux en h, est stable et permet, par conséquent, un calcul approché de la solution du problème de Dirichlet.

Mettons le problème de Dirichlet relatif à l'équation de Poisson dans le domaine carré $D = (0 \leqslant x, y \leqslant 1)$ de frontière Γ sous la forme

$$\left.\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= \varphi(x, y), \quad 0 \leqslant x, y \leqslant 1, \\ u|_\Gamma &= \psi(s), \end{aligned}\right\} \tag{1}$$

avec s la longueur d'arc comptée le long de Γ, les fonctions $\varphi(x, y)$ et $\psi(s)$ étant données.

Désignons par D_h l'ensemble des points $(x_m, y_n) = (mh, nh)$ du réseau $h = 1/M$, M entier, tombant dans l'intérieur du carré ou sur sa frontière. Les points de D_h strictement intérieurs à D seront pris pour points intérieurs du réseau carré D_h ; notons D_h^0 l'ensemble de ces points. Les points de D_h situés sur la frontière Γ de D seront pris pour points frontières de D_h et leur ensemble sera désigné par Γ_h. Ecrivons le schéma aux différences (13) du §24

$$L_h u^{(h)} = f^{(h)} \tag{2}$$

sous la forme

$$L_h u^{(h)} \equiv \left\{\begin{aligned} \Lambda_{xx} u_{mn} + \Lambda_{yy} u_{mn} &= \varphi(x_m, y_n), \quad (x_m, y_n) \in D_h^0, \\ u_{mn} &= \psi(s_{mn}), \quad (x_m, y_n) \in \Gamma_h, \end{aligned}\right. \tag{3}$$

où $\psi(s_{mn})$ est la valeur de la fonction $\psi(s)$ au point $(x_m, y_n) \in \Gamma_h$.

1. Approximation. Le second membre $f^{(h)}$ du schéma aux différences (2) est de la forme

$$f^{(h)} = \left\{\begin{aligned} &\varphi(x_m, y_n), \quad (x_m, y_n) \in D_h^0, \\ &\psi(s_{mn}), \quad (x_m, y_n) \in \Gamma_h. \end{aligned}\right. \tag{4}$$

Sous l'hypothèse que la solution $u(x, y)$ du problème (1) possède des dérivées

quatrièmes bornées, on établit l'égalité suivante moyennant la formule de Taylor :

$$\Lambda_h u \equiv \Lambda_{xx}u+\Lambda_{yy}u = \frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{h^2}{24}\left(\frac{\partial^4 u(x+\xi h, y)}{\partial x^4}+\frac{\partial^4 u(x, y+\eta h)}{\partial y^4}\right). \quad (5)$$

C'est pourquoi, pour la solution $u(x, y)$ du problème (1) on a

$$L_h[u]_h = \begin{cases} \varphi(x_m, y_n)+O(h^2), & (x_m, y_n) \in D_h^0, \\ \psi(s_{mn})+0, & (x_m, y_n) \in \Gamma_h. \end{cases} \quad (6)$$

Le résidu $\delta f^{(h)}$ apparu par substitution de $[u]_h$ dans le premier membre du schéma aux différences (2) est donc de la forme

$$\delta f^{(h)} = \begin{cases} O(h^2), & (x_m, y_n) \in D_h^0, \\ 0, & (x_m, y_n) \in \Gamma_h. \end{cases} \quad (7)$$

Munissons l'espace F_h d'éléments de la forme

$$f^{(h)} = \begin{cases} \varphi_{mn}, & (x_m, y_n) \in D_h^0, \\ \psi_{mn}, & (x_m, y_n) \in \Gamma_h, \end{cases}$$

de la norme

$$\|f^{(h)}\|_{F_h} = \max_{(mh,\, nh) \in D_h^0} |\varphi_{mn}| + \max_{(mh,\, nh) \in \Gamma_h} |\psi_{mn}|. \quad (8)$$

Alors

$$\|\delta f^{(h)}\|_{F_h} = O(h^2).$$

L'approximation du problème de Dirichlet (1) par le problème aux limites aux différences (3) est donc d'ordre 2 en h.

2. Stabilité. Définissons une norme dans l'espace U_h de fonctions définies sur le réseau D_h en posant

$$\|u^{(h)}\|_{U_h} = \max_{(mh,\, nh) \in D_h} |u_{mn}|. \quad (9)$$

Conformément à la définition de la stabilité, celle du schéma aux différences (3) se démontre en établissant l'unicité de la solution de (2) pour $f^{(h)}$ arbitraire (cette propriété ne dépend pas du choix des normes) et une estimation de la forme

$$\|u^{(h)}\|_{U_h} \leqslant c\,\|f^{(h)}\|_{F_h}, \quad (10)$$

où c est indépendant de h et de $f^{(h)}$.

LEMME 1. *Etant donné une fonction $v^{(h)} = \{v_{mn}\}$ définie sur le réseau D_h et vérifiant aux points intérieurs $(x_m, y_n) = (mh, nh) \in D_h^0$ la condition*

$$\Lambda_h v^{(h)}|_{(mh,\, nh)} \geqslant 0, \quad (mh, nh) \in D_h^0, \quad (11)$$

$v^{(h)}$ atteint sa plus grande valeur sur D_h au moins en un point de Γ_h.

DÉMONSTRATION. Procédons par l'absurde et choisissons de tous les points de D_h en lesquels $v^{(h)}$ atteint sa plus grande valeur un point (x_m, y_n)

de plus grande abscisse. Par hypothèse (x_m, y_n) est un point intérieur et v_{mn} est strictement supérieur à $v_{m+1,n}$. On a au point (mh, nh)

$$\Lambda_h v^{(h)}|_{(mh,\,nh)} \equiv$$
$$\equiv \frac{(v_{m+1,n}-v_{mn})+(v_{m,n+1}-v_{mn})+(v_{m-1,n}-v_{mn})+(v_{m,n-1}-v_{mn})}{h^2} < 0$$

du moment que la première parenthèse du nominateur est négative et les autres sont non positives : contradiction avec (11).

LEMME 2. *Etant donné une fonction $v^{(h)} = \{v_{mn}\}$ définie sur le réseau D_h et vérifiant aux points intérieurs $(mh, nh) \in D_h^0$ la condition*

$$\Lambda_h v^{(h)}|_{(mh,\,nh)} \leqslant 0, \quad (mh, nh) \in D_h^0, \tag{12}$$

$v^{(h)}$ atteint sa plus petite valeur sur D_h au moins en un point de la frontière.

La démonstration est analogue à celle du lemme 1.

THÉORÈME (PRINCIPE DU MAXIMUM). *Chaque solution de l'équation aux différences*

$$\Lambda_h v^{(h)}|_{(mh,\,nh)} = 0, \quad (mh, nh) \in D_h^0, \tag{13}$$

atteint sa plus grande et plus petite valeur en certains points de Γ_h.

On le démontre en réunissant les affirmations des deux lemmes précédents.

La propriété des solutions de l'équation aux différences (13) est analogue à la propriété des solutions $v(x, y)$ de l'équation de Laplace $v_{xx}+v_{yy} = 0$ de prendre la plus grande et plus petite valeur sur la frontière de leur domaine de définition.

Il résulte du principe du maximum que le problème

$$L_h u^{(h)} = \begin{cases} \Lambda_h u^{(h)}|_{(mh,\,nh)} = 0, & (mh, nh) \in D_h^0, \\ u^{(h)}|_{(mh,\,nh)} = 0, & (mh, nh) \in \Gamma_h, \end{cases} \tag{14}$$

n'admet que la solution nulle $u^{(h)} = 0$, vu que $u^{(h)}$ atteint sa plus grande et plus petite valeur en des points de Γ_h où $u_{mn} = 0$. Par conséquent, le déterminant du système d'équations linéaires (3) est non nul et le problème aux limites aux différences (2) admet une solution unique pour un second membre $f^{(h)}$ quelconque.

Passons à la démonstration de l'estimation (10).

En vertu de la formule (5), un polynôme $P(x, y)$ de deuxième (voire de troisième) degré

$$P(x, y) = ax^2+bxy+cy^2+dx+ey+f$$

vérifie l'égalité

$$\Lambda_h P = \frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2} \tag{15}$$

du moment que les dérivées quatrièmes de $P(x, y)$ entrant dans l'expression du reste de la formule (5) deviennent nulles.

Utilisant les fonctions φ_{mn} et ψ_{mn} du second membre du système (3) et fixant $R > \sqrt{2}$, construisons la fonction auxiliaire

$$P^{(h)}(x, y) = \frac{1}{4}[R^2-(x^2+y^2)] \max_{(mh,\, nh) \in D_h^0} |\varphi_{mn}| + \max_{(mh,\, nh) \in \Gamma_h} |\psi_{mn}|,$$

que nous ne considérerons qu'aux points du réseau D_h. Ce fait est traduit par l'indice h dans la notation $P^{(h)}(x, y)$. Vu (15) on a partout aux points de D_h^0

$$\Lambda_h P^{(h)}\Big|_{\substack{x=mh,\\ y=nh}} \equiv -\max_{(rh,\, sh) \in D_h^0} |\varphi_{rs}|, \quad (mh, nh) \in D_h^0,$$

et la différence de la solution $u^{(h)}$ du problème (3) et de la fonction $P^{(h)}$ satisfait donc en ces points aux égalités

$$\Lambda_h(u^{(h)}-P^{(h)}) = \Lambda_h u^{(h)} - \Lambda_h P^{(h)} = \varphi_{mn} + \max_{r,\, s} |\varphi_{rs}| \geqslant 0.$$

Selon le lemme 1, la différence $u^{(h)}-P^{(h)}$ prend sa plus grande valeur sur la frontière Γ_h. Or sur celle-ci cette différence

$$u^{(h)}|_{\Gamma_h} - P^{(h)}|_{\Gamma_h} = [\psi_{mn} - P^{(h)}]_{mn} =$$

$$= \Big[\psi_{mn} - \max_{(rh,\, sh) \in \Gamma_h} |\psi_{rs}|\Big] + \frac{1}{4}[(x^2+y^2)-R^2] \max_{(rh,\, sh) \in D_h^0} |\varphi_{rs}|$$

est non positive parce que partout dans le carré D on a $x^2+y^2 \leqslant R^2$ et les deux derniers crochets sont non positifs. Du moment que la plus grande valeur de $u^{(h)}-P^{(h)}$ est non positive,

$$u^{(h)} - P^{(h)}|_{(mh,\, nh)} \leqslant 0 \quad \text{ou} \quad u^{(h)} \leqslant P^{(h)}$$

partout sur D_h. De même, pour la fonction $u^{(h)}+P^{(h)}$

$$\Lambda_h(u^{(h)}+P^{(h)}) \leqslant 0$$

aux points de D_h^0, et aux points de Γ_h cette somme est non négative. En vertu du lemme 2, partout sur D_h

$$u^{(h)} + P^{(h)} \geqslant 0 \quad \text{ou} \quad -P^{(h)} \leqslant u^{(h)}.$$

Ainsi, on a partout sur D_h

$$|u_{mn}| \leqslant |P_{mn}^{(h)}| \leqslant \frac{1}{4} R^2 \max_{(rh,\, sh) \in D_h^0} |\varphi_{rs}| + \max_{(rh,\, sh) \in \Gamma_h} |\psi_{rs}|.$$

D'où l'inégalité (10) :

$$\max |u_{mn}| = \|u^{(h)}\|_{U_h} \leqslant c\, \|f^{(h)}\|_{F_h} = c\Big(\max_{(rh,\, sh) \in D_h^0} |\varphi_{rs}| + \max_{(rh,\, sh) \in \Gamma_h} |\psi_{rs}|\Big),$$

avec

$$c = \max\Big[1, \frac{R^2}{4}\Big],$$

et la stabilité est démontrée.

S'agissant du problème de Dirichlet pour l'équation elliptique à coefficients variables

$$\frac{\partial}{\partial x}\left[k_1(x, y)\frac{\partial u}{\partial x}\right]+\frac{\partial}{\partial y}\left[k_2(x, y)\frac{\partial u}{\partial y}\right] = \varphi(x, y), \quad (x, y) \in D,$$
$$u|_\Gamma = \psi(s),$$

où $k_1(x, y)$ et $k_2(x, y)$ sont deux fonctions régulières positives dans le rectangle D, le schéma aux différences se constitue de façon analogue. En remplaçant aux points intérieurs de D_h^0 $\frac{\partial}{\partial x}\left(k_1\frac{\partial u}{\partial x}\right)$ et $\frac{\partial}{\partial y}\left(k_2\frac{\partial u}{\partial y}\right)$ par des rapports aux différences moyennant les formules approchées

$$\frac{\partial}{\partial x}\left[k_1(x, y)\frac{\partial u(x, y)}{\partial x}\right] \approx \tilde{\Lambda}_{xx}u^{(h)} \equiv \frac{1}{h}\left[k_1\left(x+\frac{h}{2}, y\right)\frac{u(x+h, y)-u(x, y)}{h}-\right.$$
$$\left.-k_1\left(x-\frac{h}{2}, y\right)\frac{u(x, y)-u(x-h, y)}{h}\right],$$
$$\frac{\partial}{\partial y}\left[k_2(x, y)\frac{\partial u(x, y)}{\partial y}\right] \approx \tilde{\Lambda}_{yy}u^{(h)} \equiv \frac{1}{h}\left[k_2\left(x, y+\frac{h}{2}\right)\frac{u(x, y+h)-u(x, y)}{h}-\right.$$
$$\left.-k_2\left(x, y-\frac{h}{2}\right)\frac{u(x, y)-u(x, y-h)}{h}\right]$$

nous aboutissons à un schéma aux différences (2) de la forme

$$L_h u^{(h)} = \begin{cases} \tilde{\Lambda}_{xx}^{(h)}u^{(h)}+\tilde{\Lambda}_{yy}^{(h)}u^{(h)} = \varphi(mh, nh), & (mh, nh) \in D_h^0, \\ u|_{\Gamma_h} = \psi(s_{mn}), & (mh, nh) \in \Gamma_h. \end{cases}$$

En utilisant la formule de Taylor on se convainc que l'approximation est d'ordre deux. On pourrait démontrer la stabilité du schéma construit à force de surmonter certaines difficultés complémentaires absentes de l'exemple considéré.

Dans la pratique, on se borne généralement, pour tel ou tel problème concret, à une justification de principe sur des problèmes modèles du type ci-dessus. Des appréciations concrètes sur les erreurs découlent en général non pas des évaluations théoriques, mais de la comparaison des résultats du calcul fait avec des réseaux à h différent.

Une fois que le problème aux limites aux différences approchant le problème différentiel est construit, il faut indiquer un procédé de résolution pas trop laborieux. Car, pour un h petit, le problème (2) constitue un système d'équations scalaires d'ordre très élevé. Dans l'exemple analysé la résolution des équations aux différences est un problème intéressant et ardu dont l'examen sera fait aux §§35, 36.

Exercices

1. Démontrer que si en tous les points intérieurs du domaine D_h la fonction $u^{(h)}$ satisfait à l'équation

$$\Lambda_h u^{(h)}|_{(mh,\, nh)} = 0, \quad m, n = 1, 2, \ldots, M-1, \quad Mh = 1,$$

alors ou bien elle prend partout sur D_h une même valeur, ou bien elle n'atteint sa plus grande et sa plus petite valeur en aucun point intérieur de D_h (PRINCIPE RENFORCÉ DU MAXIMUM).

2. Si $\Lambda_h u^{(h)} \geqslant 0$ en tous les points intérieurs du domaine D_h et que cette inégalité soit stricte en un point au moins, alors $u^{(h)}$ ne réalise sa plus grande valeur en aucun point intérieur.

3. Soit le schéma aux différences $L_h u^{(h)} = f^{(h)}$ de la forme

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_{m+1,\,n}+u_{m,\,n+1}+u_{m-1,\,n}+u_{m,\,n-1}-4u_{mn}}{h^2} = \varphi(mh, nh), & (mh, nh) \in D_h^0, \\ u_{mn} = \psi_1(s_{mn}) \quad \text{sur} \quad \Gamma_h^{(1)}, \\ \dfrac{u_{1,\,n}-u_{0,\,n}}{h} = \psi_2(nh), \quad n = 1, \ldots, M-1. \end{cases}$$

Il approche le problème (fig. 43)

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2} = \varphi(x, y) \qquad (x, y) \in D,$$

$$u(x, y)|_{\Gamma^{(1)}} = \psi_1(s), \qquad (x, y) \in \Gamma^{(1)},$$

$$\left.\frac{\partial u}{\partial n}\right|_{\Gamma^{(2)}} = \psi_2(s), \qquad (x, y) \in \Gamma^{(2)}.$$

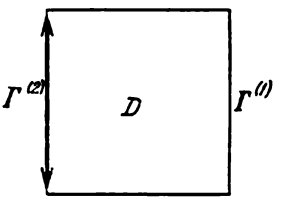


Fig. 43

a) Démontrer l'unicité de la solution du problème $L_h u^{(h)} = f^{(h)}$ quels que soient $\varphi(mh, nh)$, $\psi_1(s_{mn})$, $\psi_2(nh)$;

b) démontrer que si $\varphi(mh, nh)$ est non négatif et $\psi_1(s_{mn})$ et $\psi_2(nh)$ sont non positifs, alors $u^{(h)}$ est non positif ;

c) démontrer que, quels que soient φ, ψ_1, ψ_2, on a une estimation de la forme

$$\max_{m,\,n} |u_{mn}| \leqslant c\Big(\max_{m,\,n} |\varphi_{mn}| + \max_{(mh,\,nh) \in \Gamma_1^{(h)}} |\psi_1(s_{mn})| + \max_{n} |\psi_2(nh)|\Big),$$

avec c une constante indépendante de h. Calculer c.

§35. Méthode de stationnarisation

1. Idée de la méthode. Pour calculer les solutions d'un grand nombre de problèmes stationnaires de la physique mathématique qui décrivent les états d'équilibre, on considère ces derniers comme résultat de la stationnarisation d'un processus évoluant dans le temps dont le calcul s'avère souvent plus simple que le calcul direct de l'état d'équilibre.

Illustrons ce que nous venons de dire par l'algorithme de calcul de la solution du problème de Dirichlet aux différences

$$\left.\begin{aligned} \Lambda_{xx}u_{mn}+\Lambda_{yy}u_{mn} &= \varphi(x_m, y_n), \quad m, n = 1, 2, \ldots, M-1, \\ u_{mn}|_{\Gamma} &= \psi(s_{mn}) \end{aligned}\right\} \tag{1}$$

qui approche le problème de Dirichlet différentiel

$$\left.\begin{aligned} \frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2} &= \varphi(x, y), \quad 0 \leqslant x, y \leqslant 1, \\ u_{|\Gamma} &= \psi(s). \end{aligned}\right\} \tag{2}$$

Dans le cas du problème (1) on procède à une analyse théorique de divers algorithmes de stationnarisation à l'aide des séries finies de Fourier. Notons

de suite qu'on aborde des problèmes elliptiques aux différences analogues à (1) par des méthodes itératives sensiblement plus puissantes dont certaines seront exposées aux §§36, 37.

Pour des M tant soit peu grands, les techniques directes de résolution du problème (1) qui s'étendent au cas des coefficients variables et des domaines de frontière curviligne, par exemple la méthode d'élimination de Gauss, deviennent d'un maniement difficile et ne s'appliquent plus.

Commençons par des considérations suggestives. La solution $u(x, y)$ du problème (2) peut être comprise comme la température indépendante du temps au point (x, y) d'une plaque en équilibre thermique. Les fonctions $\varphi(x, y)$ et $\psi(s)$ signifient alors la répartition de sources de chaleur et la température à la frontière respectivement.

Considérons le problème instationnaire auxiliaire de la propagation de la chaleur

$$\left.\begin{aligned} \frac{\partial U}{\partial t} &= \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} - \varphi(x, y), \\ U|_\Gamma &= \psi(s), \\ U(x, y, 0) &= \psi_0(x, y), \end{aligned}\right\} \tag{3}$$

φ et ψ étant les mêmes que dans le problème (2) et $\psi_0(x, y)$ arbitraire.

Puisque les sources de chaleur $\varphi(x, y)$ et la température à la frontière $\psi(s)$ sont indépendantes du temps, on s'attend naturellement que la solution $U(x, y, t)$ variera avec le temps de plus en plus lentement et que la répartition de températures $U(x, y, t)$ deviendra à la limite pour $t \to \infty$ la répartition d'équilibre de températures $u(x, y)$ décrite par le problème (2). Aussi, au lieu du problème stationnaire (2), peut-on résoudre le problème instationnaire (3) jusqu'à l'époque t où sa solution cesse de varier dans les limites de précision voulue. Telle est l'idée de la méthode de résolution des problèmes stationnaires *par stationnarisation*.

Compte tenu de ce que nous venons de dire, abordons, à la place de (2), le problème (3) et considérons et comparons, au lieu du schéma aux différences (1) pour le problème (2), trois schémas différents pour le problème (3).

Ces trois schémas sont notamment : le plus simple schéma explicite

$$\left.\begin{aligned} \frac{u_{mn}^{p+1} - u_{mn}^{p}}{\tau} &= \Lambda_{xx} u_{mn}^{p} + \Lambda_{yy} u_{mn}^{p} - \varphi(x_m, y_n), \\ u_{mn}^{p+1}|_\Gamma &= \psi(s_{mn}), \\ u_{mn}^{0} &= \psi_0(x_m, y_n), \end{aligned}\right\} \tag{4}$$

le plus simple schéma implicite

$$\left.\begin{aligned} \frac{u_{mn}^{p+1} - u_{mn}^{p}}{\tau} &= \Lambda_{xx} u_{mn}^{p+1} + \Lambda_{yy} u_{mn}^{p+1} - \varphi(x_m, y_n), \\ u_{mn}^{p+1}|_\Gamma &= \psi(s_{mn}), \\ u_{mn}^{0} &= \psi_0(x_m, y_n) \end{aligned}\right\} \tag{5}$$

et le schéma de directions alternées (12) du §31

$$\left.\begin{aligned}\frac{\tilde{u}_{mn}-u^p_{mn}}{\tau} &= \frac{1}{2}\,[\Lambda_{xx}\tilde{u}_{mn}+\Lambda_{yy}u^p_{mn}-\varphi(x_m, y_n)],\\ \frac{u^{p+1}_{mn}-\tilde{u}_{mn}}{\tau} &= \frac{1}{2}\,[\Lambda_{xx}\tilde{u}_{mn}+\Lambda_{yy}u^{p+1}_{mn}-\varphi(x_m, y_n)],\\ u^{p+1}_{mn}|_\Gamma &= \tilde{u}_{mn}|_\Gamma = \psi(s_{mn}),\\ u^0_{mn} &= \psi_0(x_m, y_n).\end{aligned}\right\} \tag{6}$$

$\psi_0(x_m, y_n)$ est supposé tel qu'on ait sur la frontière

$$\psi_0|_\Gamma = \psi(s_{mn}). \tag{7}$$

Dans le cas du schéma (4), $u^{p+1} = \{u^{p+1}_{mn}\}$ se calcule en fonction de $u^p = \{u^p_{mn}\}$ connu d'après des formules explicites.

Pour trouver $u^{p+1} = \{u^{p+1}_{mn}\}$ ($u^p = \{u^p_{mn}\}$ étant déjà calculé) selon le schéma (5), il faut résoudre le problème

$$\left.\begin{aligned}\Lambda_{xx}u^{p+1}_{mn}+\Lambda_{yy}u^{p+1}_{mn}-\frac{u^{p+1}_{mn}}{\tau} &= \varphi(x_m, y_n)-\frac{u^p_{mn}}{\tau},\\ u^{p+1}_{mn}|_\Gamma &= \psi(s_{mn}),\end{aligned}\right\}$$

qui n'est nullement plus simple que le problème de départ (1). Le plus simple schéma implicite est donc à rejeter. Enfin, le calcul de $u^{p+1} = \{u^{p+1}_{mn}\}$ en fonction de $u^p = \{u^p_{mn}\}$ connus par le schéma (6) s'effectue par balayage dans la direction de l'axe Ox (pour obtenir les solutions $\{\tilde{u}_{mn}\}$ des problèmes unidimensionnels pour tout n fixe), puis par balayage dans la direction de l'axe Oy (on trouve ainsi les solutions $\{u^{p+1}_{mn}\}$ des problèmes unidimensionnels pour tout m fixe).

Le nombre d'opérations arithmétiques est en l'occurrence proportionnel au nombre d'inconnues. Pour chacun des deux schémas (4), (6) laissés en vue de leur étude ultérieure, considérons la différence

$$\varepsilon^p_{mn} \equiv u^p_{mn}-u_{mn} \tag{8}$$

entre la fonction $u^p = \{u^p_{mn}\}$ définie sur le réseau et la solution exacte $u = \{u_{mn}\}$ du problème (1), dont l'existence vient d'être démontrée (§34).

Etablissons les conditions dans lesquelles l'erreur ε^p_{mn} sur la solution u^p_{mn} du problème instationnaire tend vers zéro avec la croissance de p, ainsi que le caractère de cette tendance : choisissons un τ optimal et apprécions le volume de calcul nécessaire pour diminuer d'un nombre donné de fois la norme de l'erreur initiale

$$\varepsilon^0_{mn} = \psi_0(x_m, y_n)-u_{mn}.$$

2. Analyse d'un schéma de stationnarisation explicite. La solution $\{u_{mn}\}$

du problème (1) vérifie évidemment les équations

$$\left.\begin{aligned} \frac{u_{mn}-u_{mn}}{\tau} &= \Lambda_{xx}u_{mn}+\Lambda_{yy}u_{mn}-\varphi(x_m, y_n), \\ u_{mn}|_\Gamma &= \psi(s_{mn}), \\ u_{mn} &= u_{mn}. \end{aligned}\right\}$$

Retranchons terme à terme ces égalités de (4) ; on arrive au problème aux différences suivant pour l'erreur ε^p_{mn} :

$$\left.\begin{aligned} \frac{\varepsilon^{p+1}_{mn}-\varepsilon^p_{mn}}{\tau} &= \Lambda_{xx}\varepsilon^p_{mn}+\Lambda_{yy}\varepsilon^p_{mn}, \\ \varepsilon^{p+1}_{mn}|_\Gamma &= 0, \\ \varepsilon^0_{mn} &= \psi_0(x_m, y_n)-u_{mn}. \end{aligned}\right\} \tag{9}$$

Notons que la fonction discrète ε^p_{mn} s'annule pour chaque p, $p = 0, 1, \ldots,$ sur la frontière Γ. On peut la considérer comme élément d'un espace vectoriel de fonctions définies sur le réseau $(x_m, y_n) = (mh, nh)$, $m, n = 0, 1, \ldots, N$, et s'annulant sur Γ. La norme dans cet espace sera définie, comme au §27, par l'égalité

$$\|\varepsilon^p\| = \Big(\sum_{m,n} |\varepsilon^p_{mn}|^2\Big)^{1/2}.$$

Au §27 nous avons représenté la solution du problème (17) par la série finie de Fourier. La seule différence entre ce problème et le schéma aux différences (9) pour l'erreur $\varepsilon^p = \{\varepsilon^p_{mn}\}$ ne réside que dans la désignation de la fonction inconnue. Donc

$$\varepsilon^p = \sum_{r,s} (c_{rs}\lambda^p_{rs})\psi^{(r,s)}, \tag{10}$$

avec c_{rs} les coefficients du développement de l'erreur initiale $\varepsilon^0 = \{\varepsilon^0_{mn}\}$ en série finie de Fourier et les nombres λ_{rs} définis par la formule

$$\lambda_{rs} = 1-\frac{4\tau}{h^2}\Big(\sin^2\frac{r\pi}{2M}+\sin^2\frac{s\pi}{2M}\Big). \tag{11}$$

Les nombres $c^p_{rs} \equiv c_{rs}\lambda^p_{rs}$ sont les coefficients du développement de l'erreur $\varepsilon^p = \{\varepsilon^p_{mn}\}$ en série de Fourier suivant la base orthonormale $\psi^{(r,s)}$. Aussi

$$\|\varepsilon^p\| = \Big(\sum |c_{rs}\lambda^p_{rs}|^2\Big)^{1/2}, \quad \|\varepsilon^0\| = \Big(\sum |c_{rs}|^2\Big)^{1/2}. \tag{12}$$

D'où

$$\frac{\|\varepsilon^p\|}{\|\varepsilon^0\|} \leqslant \Big\{\max_{r,s} |\lambda_{rs}|\Big\}^p, \tag{13}$$

et on peut toujours définir ε^0 de façon à obtenir l'égalité. On prend pour cela $\varepsilon^0 = \psi^{(r',s')}$, (r', s') étant le couple de numéros tel que

$$\max_{r,s} |\lambda_{rs}| = |\lambda_{r',s'}|.$$

Ainsi, pour que $\|\varepsilon^p\|/\|\varepsilon^0\|$ tende vers zéro avec $p \to \infty$, il faut avoir

$$\max_{r,s} |\lambda_{rs}| < 1.$$

L'erreur décroît le plus vite pour un τ tel que $\max\limits_{r,s} |\lambda_{rs}|$ prenne la plus petite valeur possible. La formule (11) permet de trouver les points extrêmes λ_{rs} :

$$\lambda_{\text{gauche}} = 1 - \frac{8\tau}{h^2} \cos^2 \frac{\pi}{2M},$$

$$\lambda_{\text{droite}} = 1 - \frac{8\tau}{h^2} \sin^2 \frac{\pi}{2M}$$

(fig. 44). Avec une augmentation de τ à partir de $\tau = 0$, ces deux points se

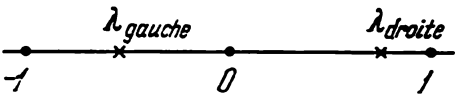


Fig. 44

déplacent à gauche. Une fois atteinte une valeur de τ telle que ces points soient symétriques par rapport à $\lambda = 0$,

$$-\lambda_{\text{gauche}} = \lambda_{\text{droite}}, \tag{14}$$

on ne gagne rien à augmenter τ. En effet, si l'on continue ce processus, λ_{droite} s'approchera toujours de zéro, mais λ_{gauche} qui lui sera supérieur en module, $\max |\lambda_{rs}| = -\lambda_{\text{gauche}}$, s'en éloignera.

Pour un τ tel que $\lambda_{\text{gauche}} = -1$ et pour ses valeurs plus grandes, l'erreur ε^p ne tend en général pas vers zéro.

Ainsi, la valeur optimale $\tau = h^2/4$ est trouvée à partir de la condition (14). Ceci étant,

$$\max_{r,s} |\lambda_{rs}| = 1 - 2 \sin^2 \frac{\pi}{2M}.$$

La norme de l'erreur initiale $\varepsilon^0 = \{\varepsilon^0_{mn}\}$ diminue donc d'un nombre donné e de fois au bout d'un nombre p de pas du processus itératif (4) tel que

$$\left(1 - 2 \sin^2 \frac{\pi}{2M}\right)^p < e^{-1}.$$

D'où

$$p \geqslant -\frac{1}{\ln\left(1 - 2 \sin^2 \frac{\pi}{2M}\right)} \approx \frac{2M^2}{\pi^2}.$$

Quel est le nombre d'opérations arithmétiques nécessaire pour diminuer l'erreur de e fois? Chaque passage de u^p à u^{p+1} exige cM^2 de ces opérations et il en faut au total $cpM^2 = O(M^4)$.

3. Schéma de directions alternées. Voyons maintenant comment se comporte l'erreur $\varepsilon^p = \{\varepsilon^p_{mn}\}$ dans le cas du schéma (6).

En raisonnant comme dans l'exemple précédent on se convainc que l'erreur ε^p satisfait alors au problème aux limites aux différences

$$\left.\begin{aligned} 2\frac{\tilde{\varepsilon}_{mn}-\varepsilon^p_{mn}}{\tau} &= \Lambda_{xx}\tilde{\varepsilon}_{mn}+\Lambda_{yy}\varepsilon^p_{mn}, \\ 2\frac{\varepsilon^{p+1}_{mn}-\tilde{\varepsilon}_{mn}}{\tau} &= \Lambda_{xx}\tilde{\varepsilon}_{mn}+\Lambda_{yy}\varepsilon^{p+1}_{mn}, \\ \tilde{\varepsilon}_{mn}|_\Gamma &= \varepsilon^p_{mn}|_\Gamma = 0, \\ \varepsilon^0_{mn} &= \psi_0(x_m, y_n)-u_{mn}. \end{aligned}\right\} \tag{15}$$

Au §27 nous avons déjà obtenu sa solution sous forme de série finie de Fourier. Comme dans le cas du problème (9), elle est de la forme (10) :

$$\varepsilon^p = \sum (c_{rs}\lambda^p_{rs})\psi^{(r,s)},$$

où c_{rs} sont les coefficients du développement de l'erreur initiale

$$\varepsilon^0 = \sum c_{rs}\psi^{(r,s)}$$

en série finie de Fourier, mais les nombres λ_{rs}, facteurs de l'harmonique $\psi^{(r,s)}$ dans le passage $\varepsilon^p \to \varepsilon^{p+1}$, sont différents, à savoir :

$$\lambda_{rs} = \frac{\left(1-2\tau M^2 \sin^2 \frac{\pi r}{2M}\right)\left(1-2\tau M^2 \sin^2 \frac{\pi s}{2M}\right)}{\left(1+2\tau M^2 \sin^2 \frac{\pi r}{2M}\right)\left(1+2\tau M^2 \sin^2 \frac{\pi s}{2M}\right)}. \tag{16}$$

Comme pour le schéma (4) dont nous avons analysé la convergence, nous avons l'inégalité (13) :

$$\frac{\|\varepsilon^p\|}{\|\varepsilon^0\|} \leqslant \left\{\max_{r,s} |\lambda_{rs}|\right\}^p,$$

l'égalité n'étant réalisée que pour $\varepsilon^0 = \{\varepsilon^0_{mn}\}$ donné de façon spéciale.

L'expression (16) de λ_{rs} implique $|\lambda_{rs}| < 1$ quel que soit τ et, par conséquent, $\|\varepsilon^p\| \to 0$. Ensuite, $\lambda_{rs} = \lambda_r \cdot \lambda_s$, où

$$\lambda_k = \frac{1-2\tau M^2 \sin^2 \frac{\pi k}{2M}}{1+2\tau M^2 \sin^2 \frac{\pi k}{2M}}, \quad k = 1, 2, \ldots, M-1.$$

C'est pourquoi $\max\limits_{r,s} |\lambda_{rs}|$ est atteint pour $r = s = r'$, r' étant le numéro pour lequel la quantité $|\lambda_{r'}|$ est maximale. Il est évident que la fonction $y = (1-x)/(1+x)$ est monotone. Donc

$$\lambda_s = \frac{1-2\tau M^2 \sin^2 \frac{\pi s}{2M}}{1+2\tau M^2 \sin^2 \frac{\pi s}{2M}}$$

se situe entre les points

$$\lambda_{\text{gauche}} = \frac{1-2\tau M^2\cos^2\dfrac{\pi}{2M}}{1+2\tau M^2\cos^2\dfrac{\pi}{2M}}$$

et

$$\lambda_{\text{droite}} = \frac{1-2\tau M^2\sin^2\dfrac{\pi}{2M}}{1+2\tau M^2\sin^2\dfrac{\pi}{2M}}$$

de la droite réelle. Le pas τ augmentant, les points λ_{gauche} et λ_{droite} se déplacent à gauche, et $\max\limits_s |\lambda_s|$ réalise sa valeur minimale pour un τ tel que $-\lambda_{\text{gauche}} = \lambda_{\text{droite}}$, i.e. pour $\tau \approx \dfrac{1}{\sqrt{2}\pi M}$. Dans ce cas

$$\max|\lambda_{rs}| = 1-\frac{\sqrt{2}\pi}{M}+O\left(\frac{1}{M^2}\right).$$

Pour diminuer la norme de l'erreur $\|\varepsilon^p\|$ *d'un nombre donné e de fois par rapport à la norme de l'erreur initiale* $\|\varepsilon^0\|$, *le nombre p de pas doit être trouvé à partir de la condition* $\left(1-\dfrac{\pi\sqrt{2}}{M}\right)^p \leqslant e^{-1}$, d'où

$$p \approx \frac{M}{\pi\sqrt{2}} = O(M).$$

Chaque passage de u^p à u^{p+1} demande cM^2 opérations arithmétiques, ce qui fait que $O(M^3)$ opérations arithmétiques sont au total nécessaires pour diminuer l'erreur de e fois et $O(M^3 \ln k)$ pour la diminuer d'un nombre donné k de fois.

Nous voyons que, lorsque M est grand, le second processus de stationnarisation (celui qui utilise le schéma de directions alternées) conduit à une diminution de l'erreur d'un nombre donné de fois qui coûte moins en opérations arithmétiques que celle obtenue par la méthode utilisant le plus simple schéma aux différences explicite (4) : pour des M suffisamment importants (pour un réseau fin), le schéma de directions alternées s'avère plus avantageux.

4. Choix de la précision. Faisons une remarque sur la précision avec laquelle il faut résoudre le problème (1) par une méthode de stationnarisation ou par une autre technique itérative donnant les approximations successives $u^1, u^2, \ldots, u^p$. Le problème aux différences (1) approche le problème (2) pour la solution régulière $u(x, y)$ à l'ordre $h^2 = 1/M^2$. Aussi la différence entre la solution exacte $u^{(h)}$ de (1) et la table cherchée $[u]_h$ est une quantité de l'ordre de $1/M^2$. On n'a donc aucune raison de calculer $u^{(h)}$ avec une précision plus grande. Si l'approximation d'ordre zéro $u^0 = \psi_0$ est supposée donnée avec une erreur de l'ordre de 1, le nombre k de fois dont on veut diminuer la norme de l'erreur doit être de l'ordre de M^2.

Vouloir diminuer l'erreur initiale de plus de $O(M^2)$ fois équivaudrait à un effort de calcul fait en pure perte.

Dans le calcul sur un réseau fixe réel on opère par itérations tant que les approximations successives u^p, u^{p+1}, ... cessent de varier à la précision désirée.

5. Limites d'applicabilité des méthodes. Le schéma aux différences (4) et l'analyse ci-dessus de la diminution de l'erreur s'étendent aux schémas aux différences approchant d'autres problèmes aux limites pour l'équation elliptique à coefficients variables dans un domaine de frontière curviligne. Il importe seulement que l'opérateur $\tilde{\Lambda}_h$ analogue à l'opérateur $-\Lambda_h \equiv \equiv -(\Lambda_{xx}+\Lambda_{yy})$ dans le schéma (1) qui agit sur des fonctions discrètes vérifiant les conditions aux limites homogènes soit auto-adjoint et que ses valeurs propres μ_j soient de même signe :

$$0 < \mu_{\min} < \mu_j < \mu_{\max}.$$

On utilise alors dans l'étude les séries finies de Fourier procédant non pas suivant les fonctions

$$\psi^{(r,s)} = 2\sin\frac{\pi r}{M}\sin\frac{\pi s}{M},$$

mais suivant un système orthonormal de fonctions propres de cet opérateur auto-adjoint $\tilde{\Lambda}_h$. Il est bien connu qu'un tel système de fonctions propres existe et est complet, et leur forme concrète n'est utilisée nulle part.

Le schéma aux différences (6) de directions alternées se généralise au problème de Dirichlet à coefficients variables dans un domaine limité par un contour curviligne. (L'analyse moyennant les séries finies de Fourier s'avère alors impossible.) Pour des conditions aux limites de la forme $\alpha u+\beta\frac{\partial u}{\partial n}\Big|_\Gamma = \psi$ une généralisation directe du schéma (6) ne conduit pas à la décomposition de l'algorithme en balayages unidimensionnels.

Exercices

1. Par analogie avec les schémas considérés (4) et (6) écrire les schémas explicite et implicite de résolution du problème de Dirichlet par stationnarisation

a) pour l'équation de Laplace à coefficients variables

$$\frac{\partial}{\partial x}\left[k_1(x,y)\frac{\partial u}{\partial x}\right]+\frac{\partial}{\partial y}\left[k_2(x,y)\frac{\partial u}{\partial y}\right] = 0, \quad 0 \leqslant x,\ y \leqslant 1,$$
$$u|_\Gamma = \psi(x,y)|_\Gamma;$$

b) pour l'équation quasi linéaire

$$\frac{\partial}{\partial x}\left[k_1(u)\frac{\partial u}{\partial x}\right]+\frac{\partial}{\partial y}\left[k_2(u)\frac{\partial u}{\partial y}\right] = 0, \quad 0 \leqslant x,\ y \leqslant 1,$$
$$u|_\Gamma = \psi(x,y)|_\Gamma.$$

2. Montrer que dans la méthode de directions alternées de résolution par itérations

du problème de Dirichlet aux différences

$$\left.\begin{aligned} &\Lambda_{xx}u_{mn}+\Lambda_{yy}u_{mn} = \varphi(x_m, y_n), \\ &\qquad m, n = 1, 2, \ldots, M-1; \quad Mh = 1, \\ &u_{mn}|_\Gamma = \psi(x, y)|_\Gamma \end{aligned}\right\}$$

on peut choisir le paramètre d'itération τ de façon qu'après la première itération le développement de l'erreur ε^p en série finie de Fourier ne contienne aucun harmonique $\psi^{(r,s)}$ donné à l'avance.

§36. Itérations à pas variable

1. Idée de Richardson. Le mécanisme de la convergence du plus simple schéma de stationnarisation (4) du §35

$$\left.\begin{aligned} \frac{u_{mn}^{p+1}-u_{mn}^p}{\tau} &= \Lambda_{xx}u_{mn}^p+\Lambda_{yy}u_{mn}^p-\varphi(x_m, y_n), \\ u_{mn}^{p+1}|_\Gamma &= \psi(s_{mn}), \\ u_{mn}^0 &= \psi_0(x_m, y_n) \end{aligned}\right\} \tag{1}$$

consiste, nous l'avons vu, à amortir chacun des harmoniques $\psi^{(r,s)}$ suivant lesquels procède le développement de l'erreur $\varepsilon_{mn}^0 = u_{mn}-u_{mn}^0$ sur l'approximation d'ordre zéro en série finie de Fourier. Si

$$\varepsilon^p = \sum_{r,s} c_{rs}^p\psi^{(r,s)}, \tag{2}$$

les coefficients de Fourier de l'erreur sur l'approximation suivante

$$\varepsilon^{p+1} = \sum_{r,s} c_{rs}^{p+1}\psi^{(r,s)}$$

s'expriment en fonction de c_{rs}^p conformément aux formules (voir (10), (11) du §35) :

$$c_{rs}^{p+1} = (1-\tau\mu_{rs})c_{rs}^p, \quad \text{où} \quad \mu_{rs} = 4M^2\left(\sin^2\frac{\pi r}{2M}+\sin^2\frac{\pi s}{2M}\right). \tag{3}$$

Pour τ fixe, tous les harmoniques ne s'amortissent pas également vite. Le processus est le plus rapide pour les harmoniques $\psi^{(r,s)}$ dont le facteur d'amortissement $\lambda_{rs} \equiv 1-\tau\mu_{rs}$ est le plus proche de zéro, i.e. pour lesquels $\mu_{rs} \approx 1/\tau$. Cela suggère l'idée de changer le paramètre τ d'un pas à l'autre pour que tous les harmoniques $\psi^{(r,s)}$ s'atténuent à tour de rôle, ce qui donne leur amortissement plus ou moins uniforme au bout de plusieurs pas.

Telle est l'idée avancée par Richardson pour les systèmes linéaires autoadjoints ayant une matrice dont toutes les valeurs propres sont de même signe.

2. Ensemble de paramètres de Tchebycheff. Le processus itératif de Richardson est défini par les formules

$$\left.\begin{aligned} &u_{mn}^{p+1} = u_{mn}^p+\tau_{p+1}[\Lambda_h u_{mn}^p-\varphi(x_m, y_n)], \quad m, n = 1, 2, \ldots, M-1, \\ &u_{mn}^{p+1}|_\Gamma = \psi(s_{mn}); \qquad \{u_{mn}^0\} \text{ donné}, \end{aligned}\right\} \tag{4}$$

le paramètre d'itération τ_{p+1} dépendant du numéro de l'itération. Richardson a proposé un ensemble de paramètres $\{\tau_p\}$ qui est acceptable sans être optimal. Exposons les résultats sur l'ensemble optimal $\{\tau_p\}$ et sur l'évaluation de la vitesse de décroissance de la norme de l'erreur $\|\varepsilon^p\|$. Il est évident qu'en vertu de la formule (3) les coefficients de Fourier c_{rs}^k de l'erreur ε^k au k-ième pas s'expriment moyennant les coefficients de Fourier c_{rs}^0 de l'erreur initiale ε^0 par les formules

$$c_{rs}^k = c_{rs}^0 \prod_{j=1}^{k} (1-\tau_j\mu_{rs}), \quad r, s = 1, 2, \ldots, M-1.$$

Introduisons la notation $Q_k(\mu)$ en posant

$$Q_k(\mu) \equiv \sum_{j=1}^{k} (1-\tau_j\mu). \tag{5}$$

Alors

$$\|\varepsilon^k\|^2 = \sum_{r,s} |c_{rs}^k|^2 = \sum_{r,s} |Q_k(\mu_{rs})c_{rs}^0|^2 \leqslant$$

$$\leqslant \max_{r,s} |Q_k(\mu_{rs})| \sum_{r,s} |c_{rs}^2|^2 = \max_{r,s} |Q_k(\mu_{rs})| \cdot \|\varepsilon^0\|^2.$$

Il est évident que pour certains ε^0 l'inégalité

$$\|\varepsilon^k\| \leqslant \max_{r,s} |Q_k(\mu_{rs})| \cdot \|\varepsilon^0\|$$

devient une égalité stricte. Les nombres μ_{rs} définis par la formule (3) sont contenus dans le segment

$$a = \mu_{\min} \leqslant \mu \leqslant \mu_{\max} = b, \tag{6}$$

où

$$\left.\begin{aligned} a &= \mu_{\min} = 8M^2 \sin^2 \frac{\pi}{2M} \approx 2\pi^2, \\ b &= \mu_{\max} = 8M^2 \cos^2 \frac{\pi}{2M} \approx 8M^2. \end{aligned}\right\} \tag{6'}$$

Nous n'allons pas profiter du fait de connaître les nombres μ_{rs}, car il s'agit là d'une circonstance fortuite n'ayant lieu que pour notre exemple. Nous n'utiliserons que notre connaissance des extrémités a et b du segment (6) qui les contient. Aussi, en définissant k, posons-nous le problème de déterminer les paramètres d'itération $\tau_1, \tau_2, \ldots, \tau_k$ de façon que de tous les polynômes $Q_k(\mu)$ de degré k tels que

$$Q(0) = 1, \tag{7}$$

le polynôme (5) s'écarte le moins possible de zéro sur le segment $a \leqslant \mu \leqslant b$:

$$Q_k^* = \max_{a<\mu<b} |Q_k(\mu)| \text{ minimal.} \tag{8}$$

Ce problème de la théorie de l'approximation des fonctions a été résolu en

1892 par A. Markov. Le polynôme cherché $Q_k(\mu) \equiv \tilde{T}_k(\mu)$ s'exprime par l'intermédiaire du polynôme de Tchebycheff (voir p. ex. V. Gontcharov, *Théorie de l'interpolation et de l'approximation des fonctions*, Moscou, 1954, éd. russe)

$$T_k(x) \equiv \cos k \arccos x \equiv \frac{1}{2}\left[(x+\sqrt{x^2-1})^k+(x-\sqrt{x^2-1})^k\right]$$

qui s'écarte le moins possible de zéro sur le segment $-1 \leqslant x \leqslant 1$ de tous les polynômes de degré k, de coefficient 1 devant x^k. Notamment, en effectuant la transformation linéaire

$$x = \frac{b+a-2\mu}{b-a} \tag{9}$$

qui change le segment $a \leqslant \mu \leqslant b$ en le segment $-1 \leqslant x \leqslant 1$ et le point $\mu = 0$ en le point $x_0 = \frac{b+a}{b-a} > 1$, on a

$$\tilde{T}_k(\mu) = \frac{T_k(x)}{T_k(x_0)} = \frac{(x+\sqrt{x^2-1})^k+(x-\sqrt{x^2-1})^k}{(x_0+\sqrt{x_0^2-1})^k+(x_0-\sqrt{x_0^2-1})^k}. \tag{10}$$

L'ensemble de paramètres d'itération $\tau_1, \tau_2, \ldots, \tau_k$ déterminant l'apparition du polynôme (10) se définit à partir de la condition que les zéros $\mu_j = 1/\tau_j$ du polynôme $\tilde{T}_k(\mu)$ se transforment par (9) en les zéros x_j du polynôme de Tchebycheff $T_k(x)$:

$$\left.\begin{aligned} \tau_j &= \frac{2}{b+a-(b-a)x_j}, \\ x_j &= \cos\frac{\pi(2j-1)}{2k}, \quad j = 1, 2, \ldots, k. \end{aligned}\right\} \tag{11}$$

Evaluons l'écart maximal Q_k^* entre le polynôme $Q_k(\mu) \equiv \tilde{T}_k(\mu)$ et le zéro sur le segment $a \leqslant \mu \leqslant b$. La théorie de l'approximation des fonctions dit que le polynôme de Tchebycheff $T_k(x)$ prend sur le segment $-1 \leqslant x \leqslant 1$ la valeur de plus grand module en $k+1$ points dont les extrémités du segment. Il résulte donc de (10)

$$Q_k^* = \frac{T_k(1)}{T_k(x_0)} = \frac{2}{(x_0+\sqrt{x_0^2-1})^k+(x_0-\sqrt{x_0^2-1})^k}. \tag{12}$$

La formule (9) donne ensuite

$$\left.\begin{aligned} x_0 &= \frac{b+a}{b-a} = \frac{1+\eta}{1-\eta} = 1+2\eta+O(\eta^2), \\ \eta &= \frac{a}{b} = \frac{\mu_{\min}}{\mu_{\max}} \approx \frac{\pi^2}{4M^2}. \end{aligned}\right\} \tag{13}$$

C'est pourquoi, pour des M grands

$$x_0 \pm \sqrt{x_0^2-1} = 1 \pm 2\sqrt{\eta}+O(\eta),$$

d'où, compte tenu de (13),

$$Q_k^* = \frac{2}{[1+2\sqrt{\eta}+O(\eta)]^k+[1-2\sqrt{\eta}+O(\eta)]^k} =$$
$$= 2:\left\{e^{k\ln(1+2\sqrt{\eta}+O(\eta))}+e^{k\ln(1-2\sqrt{\eta}+O(\eta))}\right\} \approx$$
$$\approx 2:\left\{e^{k\pi/M}+e^{-k\pi/M}\right\}.$$

La norme de l'erreur initiale ε^0 étant supposée de l'ordre de l'unité: $\|\varepsilon^0\| \approx 1$, en vertu de la remarque (n° 4 du §35) sur l'ordre raisonnable de la précision à atteindre en résolvant un problème par itérations, on doit choisir k à partir de la condition $Q_k^* \approx M^{-2}$, i.e.

$$k \approx \frac{2\ln M+\ln 2}{\pi} M \approx \frac{2\ln M+\ln 2}{2\sqrt{\eta}}. \tag{14}$$

Afin d'amortir l'erreur initiale de e fois, on doit choisir k à partir de la condition $Q_k^* \leqslant e^{-1}$, *i.e.*

$$k \approx \frac{1+\ln 2}{\pi} M \approx \frac{1+\ln 2}{2\sqrt{\eta}} = O(M). \tag{15}$$

Avec un k ainsi choisi, on peut prendre k premiers pas pour premier tour d'itération et répéter tout le tour avec un même jeu de paramètres $\tau_1, \tau_2, \ldots$ $\ldots, \tau_k$. Pour diminuer la norme de l'erreur de M^2 fois, le nombre ν de tours doit être tel que $e^{-\nu} \sim 1/M^2$, $\nu \sim 2\ln M$. Les pas élémentaires du processus itératif seront au total au bout de ν tours

$$k\nu \approx \left(\frac{1+\ln 2}{\pi} M\right) 2\ln M = O(M\ln M).$$

Ce nombre n'est supérieur qu'un nombre fini de fois, à savoir

$$2\frac{1+\ln 2}{2+\ln 2/(\ln M)} \leqslant 1+\ln 2,$$

au nombre (14) de pas élémentaires du processus itératif qui ne cycle pas. Ainsi, un tour avec $k \approx (1+\ln 2)/(2\sqrt{\eta})$ permet de simplifier les calculs sans pour autant augmenter sensiblement le nombre d'itérations.

Des tours de longueur $k \ll 1/(2\sqrt{\eta})$ sont à condamner. Par exemple, pour $k = 1$, le processus de Richardson (4) se transforme en le processus itératif simple (1) avec τ choisi de façon optimale. Le nombre de pas nécessaire pour diminuer de e fois la norme $\|\varepsilon^0\|$ de l'erreur initiale sera $\approx 2M^2/\pi^2$ (voir n° 2 du §35), soit $O(M)$ fois le nombre de pas en choisissant la longueur k du tour conformément à (15).

3. Numérotation des paramètres d'itération. Les formules (11) donnent le jeu optimal de paramètres $\tau_1, \tau_2, \ldots, \tau_k$ (pour k fixe donné). Permutons les termes de la suite $\tau_1, \tau_2, \ldots, \tau_k$ de façon à les ranger dans un certain ordre $\varkappa^{(k)} = (\varkappa_1, \varkappa_2, \ldots, \varkappa_k)$ et procédons itérativement par les formules

$$\left.\begin{aligned} u^{p+1} &= u^p+\tau_{\varkappa_{p+1}}(\Lambda_h u^p-\varphi), \\ u^{p+1}|_\Gamma &= \psi, \quad u^0 \text{ donné}. \end{aligned}\right\} \tag{16}$$

Dans une mise en œuvre idéale de l'algorithme (16) le résultat de la k-ième itération (la dernière) ne dépend pas de l'ordre choisi $\varkappa^{(k)} = (\varkappa_1, \varkappa_2, \ldots, \varkappa_k)$. Dans le calcul réel sur une machine de capacité limitée, cet ordre est de la première importance. Il détermine essentiellement, pour des k grands, la sensibilité du résultat aux erreurs d'arrondi commises aux pas intermédiaires, i.e. la stabilité numérique de l'algorithme. Avant de citer des ordres $\varkappa^{(k)} = (\varkappa_1, \varkappa_2, \ldots, \varkappa_k)$ acceptables, nous voulons indiquer l'impuissance pratique de l'algorithme de départ (4) correspondant à $\varkappa^{(k)} = (1, 2, \ldots, k)$.

Quelle est la cause de l'instabilité dans ce cas ? Supposons que l'erreur initiale ε^0 est de la forme $\varepsilon^0 = \sum c_{rs}^0 \psi^{(r,s)}$, $c_{rs}^0 \approx 1$, et que le calcul se fait exactement sans arrondi. Les coefficients de l'erreur sur la l-ième approximation $\varepsilon^l = \sum c_{rs}^l \psi^{(r,s)}$ s'expriment alors par la formule

$$c_{rs}^l = \prod_{j=1}^{l} (1-\tau_j \mu_{rs}) c_{rs}^0.$$

Suivons le comportement de c_{rs}^l avec l croissant pour $r = M-1$, $s = M-1$. On a

$$\left.\begin{aligned} \mu_{rs} &= \mu_{M-1,\,M-1} = \mu_{\max} = b \sim M^2, \\ \xi^l &\equiv \frac{c_{M-1,\,M-1}^l}{c_{M-1,\,M-1}^0} = \prod_{j=1}^{l} (1-\tau_j b). \end{aligned}\right\} \tag{17}$$

Considérons les fonctions linéaires $1-\tau_j\mu$, $j = 1, 2, \ldots, k$, dont les zéros $\mu_j = 1/\tau_j$ se définissent par les formules (11). On voit de ces dernières que

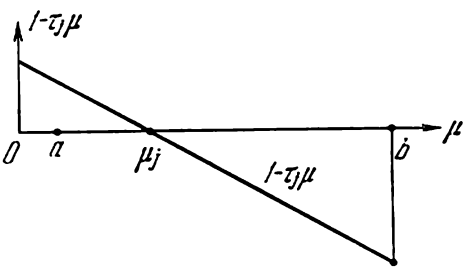


Fig. 45

lorsque $\frac{2j-1}{2k} < \frac{1}{6}$ ou $j < \frac{k+3}{6}$, on a l'inégalité $\mu_j < \frac{b}{2}$ et donc (fig. 45)

$$|1-\tau_j b| > 1 \quad \text{pour} \quad j < \frac{k+3}{6}. \tag{18}$$

Si $k \sim \frac{1}{2\sqrt{\eta}} \sim M$ et $j \sim 1$, alors $\mu_j \sim a + \frac{1}{M^2}$, $\tau_j \sim \frac{1}{a} - \frac{1}{M^2}$ et donc

$$|1-\tau_j b| \sim \frac{b}{a} \sim M^2.$$

Ainsi, la quantité ξ^l définie par (17) commence par augmenter d'environ M^2 fois par pas, puis ce processus ralentit. En vertu de (18), cette augmentation continue en tous cas tant que $l \leqslant (k+3)/6$ de sorte que pour $l \sim k/6$ la quantité $c^l_{M-1,M-1}$, de même que $\|\varepsilon^l\|$, s'avère très grande, et elle est d'autant plus grande que k l'est. Les ordres de grandeur des approximations $u^l = \{u^l_{mn}\}$ peuvent alors être tels qu'il y ait dépassement de capacité même pour des k modérés, $k \ll M$.

Supposons que tel n'est pas notre cas et qu'on continue de calculer exactement. Alors à l'entrée du pas $l = k$, la quantité $c^l_{M-1,M-1}$ diminue de nouveau, de sorte que $\xi^k \leqslant Q^*_k$.

Mais le fait est que même en l'absence de débordement pour $l \sim k/6$, il n'en reste pas moins que de petites erreurs d'arrondi relatives inévitables sont en l'occurrence grandes en valeur absolue. Il s'agit d'erreurs accidentelles si bien que dans leur développement en série finie de Fourier figureront tous les termes, y compris un terme de la forme

$$\tilde{c}_{1,1}\psi^{(1,1)},$$

$\tilde{c}_{1,1}$ n'étant pas petit en valeur absolue.

Montrons qu'aux pas suivants l'erreur $\tilde{c}_{1,1}\psi^{(1,1)}$ sur l'harmonique $\psi^{(1,1)}$ due à l'arrondi au pas $l \sim k/6$ ne s'atténue pas sensiblement et altère catastrophiquement le résultat. Sa contribution $\tilde{c}^k_{1,1}\psi^{(1,1)}$ à l'approximation u^k obtenue au k-ième (le dernier) pas s'exprime par la formule

$$\tilde{c}^k_{1,1} = \left[\prod_{j=l+1}^{k}(1-\tau_j\mu_{11})\right]\tilde{c}_{1,1} = \left[\prod_{j=l+1}^{k}(1-\tau_j a)\right]\tilde{c}_{1,1}.$$

Mais pour $j > (k+3)/6$ on a évidemment

$$\mu_j > \frac{1}{2}\left[b+a-\frac{b-a}{2}\right] > \frac{b}{4} \sim M^2.$$

Donc

$$1-\tau_j a \sim 1-\frac{1}{M^2}, \quad \left|\prod_{j=l+1}^{k}(1-\tau_j a)\right| \sim \left(1-\frac{1}{M^2}\right)^{k-1} \sim 1,$$

de sorte que $\tilde{c}^k_{1,1} \sim \tilde{c}_{1,1}$ et l'erreur d'arrondi ne s'amortit pas. Ainsi, la numérotation $\varkappa^{(k)} = (1, 2, \ldots, k)$ des paramètres est impraticable.

Si, en arrondissant au l-ième pas de (13), on fait une erreur

$$c_{rs}\psi^{(r,s)},$$

elle devient

$$\left[\prod_{j=l+1}^{k}(1-\tau_j\mu_{rs})\right]c_{rs}\psi^{(r,s)}$$

au k-ième pas. On s'appliquera donc à choisir un ordre $\varkappa^{(k)} = (\varkappa_1, \varkappa_2, \ldots, \varkappa_k)$

tel qu'on ait l'estimation

$$\max_{a<\mu<b}\left|\prod_{j=l+1}^{k}(1-\tau_j\mu)\right|<A \tag{19}$$

pour une valeur modérée de A.

Soit $c_{rs}^0\psi^{(r,s)}$ une composante de l'erreur ε^0 sur l'approximation d'ordre zéro. Au l-ième pas elle devient

$$\left[\prod_{j=1}^{l}(1-\tau_j\mu_{rs})\right]c_{rs}^0\psi^{(r,s)}.$$

Si la norme de cette fonction est grande, celle-ci donne après arrondi une contribution importante (en valeur absolue) dans tous les harmoniques. Il se peut que pour certains harmoniques cette contribution ne s'affaiblisse pas aux itérations suivantes et altère considérablement le résultat. Il est donc naturel de choisir un $\varkappa^{(k)}=(\varkappa_1, \varkappa_2, \ldots, \varkappa_k)$ tel qu'on ait l'inégalité

$$\max_{a<\mu<b}\left|\prod_{j=1}^{l}(1-\tau_j\mu)\right|<B \tag{20}$$

avec un B modéré.

Le travail de V. Lébédev et de S. Finoguénov, Zh. Vychisl Mat. i Mat. Fiz. **11**, n° 2 (1971) et le livre de A. Samarski [12] indiquent divers procédés rationnels de numérotage des paramètres et parlent de l'historique de la question. Reproduisons les résultats de Lébédev et Finoguénov. Ces auteurs supposent que k est une puissance de 2, i. e. $k=2^i$, et donnent les formules récurrentielles permettant de construire $\varkappa^{(k)}$.

Notamment, pour $i=1$

$$\varkappa^{(k)}=(1, 2).$$

Si, pour $k=2^{i-1}$, l'ordre $\varkappa^{(2^{i-1})}$ est déjà déterminé :

$$\varkappa^{(2^{i-1})}=(\varkappa_1, \varkappa_2, \ldots, \varkappa_{2^{i-1}}),$$

on pose alors

$$\varkappa^{(2i)}=(\varkappa_1, 2^i+1-\varkappa_1, \varkappa_2, 2^i+1-\varkappa_2, \ldots, 2^i+1-\varkappa_{2^{i-1}}). \tag{21}$$

En particulier, pour $i=2, 3, 4$ on obtient successivement

$$(1, 4, 2, 3)\ ; \qquad (1, 8, 4, 5, 2, 7, 3, 6)\ ;$$
$$(1, 16, 8, 9, 4, 13, 5, 12, 2, 15, 7, 10, 3, 14, 6, 11).$$

Avec l'ordre de disposition (21) des paramètres d'itération les nombres A et B des inégalités (19) et (20) peuvent être pris indépendants de k et de l.

L'algorithme d'ordonnancement des paramètres de Samarski est un peu plus compliqué, mais k n'est pas nécessairement une puissance de 2. Ce nombre peut être de la forme $k=(2j+1)2^i$. Au lieu de (19), (20) on a d'autres estimations garantissant la stabilité dans un certain sens.

4. Méthode de Douglas-Rachford. Dans le schéma de directions alternées (6) du §35 le paramètre d'itération τ sera supposé dépendre du numéro

du pas en posant

$$\frac{\tilde{u}_{mn}-u^p_{mn}}{\tau_{p+1}} = \frac{1}{2}[\Lambda_{xx}\tilde{u}_{mn}+\Lambda_{yy}u^p_{mn}-\varphi(x_m, y_n)],$$
$$\frac{u^{p+1}_{mn}-\tilde{u}_{mn}}{\tau_{p+1}} = \frac{1}{2}[\Lambda_{xx}\tilde{u}_{mn}+\Lambda_{yy}u^{p+1}_{mn}-\varphi(x_m, y_n)],$$
$$u^{p+1}_{mn}|_\Gamma = \tilde{u}_{mn}|_\Gamma = \psi(s_{mn}), \quad m, n = 1, 2, \ldots, M-1.$$

L'erreur $\varepsilon^k = u^k - u$ s'exprimera alors par

$$\varepsilon^k = \sum_{r,s=1}^{M-1} c^k_{rs}\psi^{(r,s)}, \; c^k_{rs} = \left[\prod_{j=1}^{k} \lambda_r(\tau_j)\,\lambda_s(\tau_j)\right]c^0_{rs},$$

où

$$\lambda_i(\tau) \equiv \frac{1-2\tau M^2 \sin^2\dfrac{\pi i}{2M}}{1+2\tau M^2 \sin^2\dfrac{\pi i}{2M}}, \quad i = 1, 2, \ldots, k.$$

Pour k donné on estimera optimal un choix de nombres $\tau_1, \tau_2, \ldots, \tau_k$ tel que la quantité

$$\max_{r,s}\left|\prod_{j=1}^{k} \lambda_r(\tau_j)\,\lambda_s(\tau_j)\right|$$

prenne la valeur minimale. Dans le cas où, au lieu d'utiliser les valeurs exactes de $\lambda_r(\tau)$ et $\lambda_s(\tau)$, on se borne à localiser ceux-ci, on voit apparaître un problème du type de Tchebycheff pour le produit de fonctions homographiques, problème semblable à celui relatif à des polynômes que nous avons considéré au n° 2.

La position de ce problème, comme d'ailleurs l'idée de se servir, pour résoudre l'équation de Poisson, du processus de stationnarisation avec un schéma de directions alternées, est due à Douglas, Peaceman et Rachford. Dans leur ouvrage de 1956 *(On the numerical solution of heat conduction problems in two and three space variables)* que nous allons exposer, Douglas et Rachford ont donné une résolution approchée de ce problème. Ils ont choisi les paramètres d'itération de façon que le nombre de pas du processus itératif nécessaire pour diminuer l'erreur de e fois est $O(\ln M)$ et le nombre d'opérations arithmétiques $O(M^2 \ln M)$.

Montrons qu'en nous donnant un q positif, $q < 1$, nous pouvons choisir $k = O(\ln M)$ paramètres d'itération $\tau_1, \tau_2, \ldots, \tau_k$ de manière qu'on ait les inégalités

$$|[\lambda_r(\tau_1)\,\lambda_s(\tau_1)]\,[\lambda_r(\tau_2)\,\lambda_s(\tau_2)]\ldots[\lambda_r(\tau_k)\,\lambda_s(\tau_k)]| < q, \tag{22}$$
$$r, s = 1, 2, \ldots, M-1.$$

Alors $\|\varepsilon^k\| \leqslant q \leqslant \|\varepsilon^0\|$. A condition d'effectuer les k premières itérations avec $\tau_1, \tau_2, \ldots, \tau_k$, puis les k itérations suivantes avec les mêmes paramètres, la norme de l'erreur diminue évidemment de e fois au bout d'un nombre (indépendant de M) de ces tours de $k = O(\ln M)$ itérations chacun.

Justifions l'inégalité (22) et expliquons comment choisir les paramètres $\tau_1, \tau_2, \ldots, \tau_k$. Il est évident que

$$|\lambda_i(\tau)| < 1, \; i = 1, 2, \ldots, M-1, \; \tau > 0.$$

Pour que l'inégalité (22) ait lieu, quels que soient $r, s = 1, 2, \ldots, M-1$, il suffit donc que pour tout $i = 1, 2, \ldots, M-1$, au moins l'un des k facteurs $\lambda_i(\tau_p)$, $p = 1, 2, \ldots, k$, vérifie l'inégalité

$$|\lambda_i(\tau_p)| = \left| \frac{1-2\tau_p M^2 \sin^2 \frac{\pi i}{2M}}{1+2\tau_p M^2 \sin^2 \frac{\pi i}{2M}} \right| \leqslant \sqrt{q}. \tag{23}$$

Tous les nombres $2M^2 \sin^2 \frac{\pi i}{2M}$, $i = 1, 2, \ldots, M-1$, appartiennent au segment

$$a \leqslant 0{,}5\pi^2 \leqslant \mu \leqslant 2M^2 = b^2. \tag{24}$$

Ainsi, pour avoir (22) il suffit, vu (23), que chaque valeur de μ du segment (24) satisfasse au moins pour un τ, $\tau = \tau_1, \tau_2, \ldots, \tau_k$, aux inégalités

$$-\sqrt{q} < \frac{1-\tau_p\mu}{1+\tau_p\mu} < \sqrt{q}$$

et a fortiori aux inégalités

$$-\sqrt{q} \leqslant 1-\tau_p\mu \leqslant \sqrt{q}.$$

Il faut pour cela que pour chaque μ du segment (24) on ait, pour un certain τ_p, $p = 1, 2, \ldots, k$,

$$1-\sqrt{q} \leqslant \tau_p\mu \leqslant 1+\sqrt{q}. \tag{25}$$

Définissons μ_p et τ_p par les formules suivantes :

$$\mu_p = \left(\frac{1+\sqrt{q}}{1-\sqrt{q}}\right)^{p-1} a, \quad p = 1, 2, \ldots, k,$$

$$\tau_p = \frac{1-\sqrt{q}}{\mu_p}, \qquad p = 1, 2, \ldots, k. \tag{26}$$

μ augmentant de μ_p à μ_{p+1}, le nombre $\tau_p\mu$ parcourt alors le segment (25).

Il est évident qu'en choisissant k à partir de la condition $\mu_k \geqslant b$, i.e.

$$\left.\begin{aligned} k \geqslant A \ln \frac{b}{a} + 1 &= A\left(2 \ln M + \ln \frac{4}{\pi^2}\right) + 1, \\ A &= \frac{1}{\ln \frac{1+\sqrt{q}}{1-\sqrt{q}}}, \end{aligned}\right\} \tag{27}$$

on obtient moyennant (26) la suite $\tau_1, \tau_2, \ldots, \tau_k$ qui nous intéresse.

Exercices

1. Est-il possible de choisir des paramètres d'itération $\tau_1, \tau_2, \ldots, \tau_k$ tels qu'on résolve exactement le problème de Dirichlet aux différences au bout d'un nombre fini de pas du processus itératif (4) ?

Combien d'itérations cela demande ? Peut-on généraliser ce procédé de résolution au cas où on ignore les valeurs exactes de μ_{rs} ?

2. Expliquer le mécanisme de l'instabilité numérique qui se propage en calculant par les formules (16) pour

$$\varkappa^{(k)} = (k, k-1, k-2, \ldots, 2, 1)$$

et k et M importants.

Quels harmoniques $\psi^{(r,s)}$ prédomineront dans la série de Fourier pour l'erreur ε^k résultant du calcul pour $\varkappa^{(k)} = (k, k-1, \ldots, 1)$ avec arrondi ?

3. Soit A un opérateur auto-adjoint dont les valeurs propres appartiennent au segment $0 < \mu_{\min} < \mu < \mu_{\max}$. A quelle condition doit satisfaire le nombre de conditionnement $\eta = \mu_{\min}/\mu_{\max}$ pour que le processus de Richardson destiné à la résolution de l'équation $Ax = \varphi$

$$x^{p+1} = x^p - \tau_{p+1}(Ax^p - \varphi),$$

$\tau_j = 1/\mu_j$, $\mu_{\min} < \mu_j < \mu_{\max}$, $j = 1, 2, \ldots, k$, k quelconque, étant choisi de façon arbitraire, converge et soit numériquement stable ?

4. Pourquoi dans le schéma de Douglas-Rachford, l'ordre dans lequel on utilise les paramètres $\tau_1, \tau_2, \ldots, \tau_k$ n'influe pas de façon essentielle sur la stabilité numérique du processus itératif ?

5. Sous l'hypothèse que le processus itératif de Douglas-Rachford coûte par pas en temps machine vingt fois plus que le processus de Richardson, établir moyennant les formules (15) et (27) les M pour lesquels la première méthode devient plus puissante.

§37. Méthode de Fédorenko

R. Fédorenko a proposé (Zh. Vychisl. Mat. i Mat. Fiz. **1**, n° 5 (1961)) une technique de résolution par itérations des problèmes elliptiques aux différences, qu'il a appelée méthode de relaxation. Avec cette méthode, la norme de l'erreur initiale est diminuée de e fois au bout de cM^2 opérations arithmétiques, M étant le nombre de pas dans une direction et c une constante indépendante de M. Rappelons que le coût correspondant du procédé le mieux convergent parmi ceux que nous avons examinés (et en général parmi tous les procédés connus), à savoir de la méthode de Douglas-Rachford, est de $O((\ln M)M^2)$ opérations arithmétiques.

Le domaine d'applicabilité de la méthode de Fédorenko est presque le même que celui de la plus simple méthode de stationnarisation. Une contrainte supplémentaire est la « régularité » des premières fonctions propres, condition qui est en général remplie pour les problèmes elliptiques.

Dans des exemples simples, l'économie de temps machine en comparaison d'autres méthodes itératives, les meilleures en ce qui concerne la vitesse de convergence, est parfaitement mise en évidence dès $M \approx 50$. Il ne faut pas oublier que la logique du calcul dans la méthode de relaxation est beaucoup plus compliquée que disons pour le schéma de Richardson, et le temps machine dépend donc essentiellement de la qualité du programme.

Pour la plus simple estimation de la vitesse de convergence (pour l'approximation aux différences de l'équation de Poisson dans un domaine

carré sur un réseau carré, les valeurs de la solution sur la frontière étant données) voir Fédorenko, Zh. Vychisl. Mat. i Mat. Fiz. **4**, n° 3 (1964).

Dans son travail publié dans le Zh. Vychisl. Mat. i Mat. Fiz. **6**, n° 5 (1966) N. Bakhvalov étudie la convergence de la méthode de Fédorenko et obtient le même résultat pour l'analogue aux différences du premier problème aux limites dans un rectangle pour l'équation elliptique générale à coefficients réguliers

$$a_{11}\frac{\partial^2 u}{\partial x^2}+2a_{12}\frac{\partial^2 u}{\partial x\,\partial y}+a_{22}\frac{\partial^2 u}{\partial y^2}+a_1\frac{\partial u}{\partial x}+a_2\frac{\partial u}{\partial y}+au=f_0.$$

Enfin, G. Astrakhantsev (Zh. Vychisl. Mat. i Mat. Fiz. **11**, n° 2 (1971)) a abouti à un résultat analogue pour l'approximation aux différences du troisième problème aux limites relatif à l'équation elliptique auto-adjointe dans un domaine bidimensionnel arbitraire de frontière régulière.

Les justifications étant assez lourdes, nous nous bornerons à décrire l'idée de la méthode et l'algorithme de Fédorenko en renvoyant un lecteur désireux de connaître les démonstrations aux travaux en question et à l'article de synthèse de Fédorenko dans Uspehi Mat. Nauk **28**, fasc. 2 (1973).

1. Idée de la méthode. En résolvant itérativement le problème

$$\left.\begin{aligned} \Lambda_h u_{mn}-\varphi(x_m, y_n) &= 0, \quad m, n = 1, 2, \ldots, M-1,\\ u_{mn}|_\Gamma &= \psi(s_{mn})\end{aligned}\right\} \tag{1}$$

prenons pour point de départ le simple processus de stationnarisation (4) du § 35

$$\left.\begin{aligned} u_{mn}^{p+1} &= u_{mn}^p+\tau(\Lambda_h u_{mn}^p-\varphi(x_m, y_n)), \quad m, n = 1, 2, \ldots, M-1,\\ u_{mn}^{p+1}|_\Gamma &= \psi(s_{mn}), \quad \{u_{mn}^0\} \text{ donné},\end{aligned}\right\} \tag{2}$$

qui converge fort lentement dans son ensemble, mais non uniformément pour différents harmoniques. L'erreur $\varepsilon^p = u^p-u$ s'écrit, conformément à (10) du § 35, sous forme de série finie de Fourier

$$\varepsilon^p = \sum_{r,s=1}^{M-1}[\lambda_{rs}(\tau)]^p c_{rs}^0\psi^{(r,s)}, \tag{3}$$

avec c_{rs}^0 les coefficients du développement de l'erreur $\varepsilon^0 = u^0-u$ sur l'approximation d'ordre zéro et $\lambda_{rs} = 1-4\tau M^2\left(\sin^2\frac{\pi r}{2M}+\sin^2\frac{\pi s}{2M}\right)$. Les nombres λ_{rs} se situent sur le segment $\lambda_{\text{gauche}} \leqslant \lambda \leqslant \lambda_{\text{droite}}$, où

$$\lambda_{\text{gauche}} = \lambda_{M-1,\,M-1} \approx 1-8\tau M^2,$$
$$\lambda_{\text{droite}} = \lambda_{1,\,1} \approx 1-2\pi^2\tau.$$

Posons

$$\tau = \frac{3}{16M^2}. \tag{4}$$

Si dans cette condition l'un au moins des nombres r ou s est supérieur à $M/2$, alors

$$|\lambda_{rs}| < 0{,}6. \tag{5}$$

Aussi la contribution des harmoniques $\psi^{(r,s)}$ hautes fréquences, $r > M/2$ ou $s > M/2$,

dans l'erreur (3) diminue presque de moitié à chaque pas du processus itératif pour devenir bientôt insignifiante. Après plusieurs itérations par la formule (2) l'erreur ε^p ne contient comme partie essentielle que la composante régulière (les harmoniques $\psi^{(r,s)}$, $r < M/2$, $s < M/2$) parce que les harmoniques $\psi^{(r,s)}$ basses fréquences ont pour facteurs les nombres λ_{rs} plus proches de l'unité. La contribution du premier harmonique $\psi^{(1,1)}$ diminue très lentement : pour τ choisi

$$\lambda_{1,1} \approx 1-\frac{3\pi^2}{8M^2} \ (\approx 1). \tag{6}$$

Désignons par U l'approximation u^p obtenue au cours des itérations (2) et par v l'erreur $\varepsilon^p = u^p - u = U - u$. A supposer v connu, nous aurions trouvé la solution cherchée $u = U - v$. Or nous ne connaissons rien de v sinon qu'il vérifie l'équation

$$\Lambda_h v = \xi, \qquad v|_\Gamma = 0, \tag{7}$$

où ξ, fonction connue définie sur le réseau, est le résidu obtenu par substitution de $u^p = U$ dans (1) :

$$\xi = \Lambda_h u^p - \varphi = \Lambda_h u^p - \Lambda_h u = \Lambda_h(u^p - u) = \Lambda_h v.$$

Si le problème (7) de définition du terme correctif v est plus simple que le problème de départ (1), c'est uniquement en ce sens qu'on connaît d'avance que v est une fonction régulière. Voulant définir v on considère donc de façon approchée, au lieu de (7), le même problème sur un réseau de mailles deux fois plus grandes qui (pour M pair) est une partie du réseau initial :

$$\Lambda_{2h} v^* = \xi^*, \qquad v^*|_{\Gamma^*} = 0. \tag{1*}$$

On a marqué par l'astérisque les quantités sur le réseau plus lâche. Résolvons le problème (1*) par itérations à l'aide des formules

$$\left.\begin{aligned} &(v^*_{mn})^{p+1} = (v^*_{mn})^p + \tau^*[\Lambda_{2h}(v^*_{mn})^p - \xi^*_{mn}], \quad m, n = 1, 2, \ldots, M^*-1, \\ &(v^*_{mn})^{p+1}|_{\Gamma^*} = 0, \end{aligned}\right\} \tag{2*}$$

en acceptant pour approximation d'ordre zéro $(v^*_{mn})^0 \equiv 0$. Ici $M^* = M/2$, $\tau^* = \dfrac{3}{16(M^*)^2} = 4\tau$.

Chaque pas des itérations (2*) exige quatre fois moins d'effort que celui des (2) étant donné que les points de calcul sont quatre fois moins nombreux. De plus, vu $\tau^* = 4\tau$, la composante à décroissance la plus lente de l'erreur s'amortit plus vite. Conformément à (6)

$$\lambda^*_{1,1} = 1-\frac{3\pi^2}{8(M^*)^2} = 1-4\frac{3\pi^2}{8M^2} < \lambda_{1,1},$$

et la contribution de $\psi^{(1,1)}$ s'amortit de e fois au bout d'un nombre d'itérations quatre fois plus petit. Notons V^* le résultat des itérations (2*) et interpolons-le (linéairement) sur le réseau initial. Les composantes régulières seront obtenues quasi exactement. L'erreur due à l'interpolation sera faible devant la fonction régulière à interpoler, mais son développement de Fourier contiendra tous les harmoniques (l'erreur d'interpolation est non régulière à cause des inflexions aux points d'interpolation). En outre, la composante non régulière de V^* sans rapport avec la correction cherchée donne également par interpolation une contribution aléatoire dans la composante non régulière de la fonction V obtenue par interpolation. Ainsi, la composante régulière de la différence $U - V$ est proche de celle de la solution cherchée $u = U - v$, mais la composante non régulière n'est pas très petite et elle est aléatoire.

Il faut donc faire plusieurs autres pas du processus (2) en prenant $U - V$ pour approximation initiale. Les itérations (2) réduisant d'environ deux fois par pas la composante non régulière de l'erreur, qui est due à l'interpolation, celle-là s'amortit rapidement.

2. Description de l'algorithme. L'accélération de la convergence, résultat de l'utilisation d'un réseau plus lâche et du processus itératif (2*), peut se trouver insuffisante. Pour

M grand (réseau fin), le problème (1*) sur le réseau plus lâche reste laborieux. On le résout donc par passage à un réseau deux fois moins serré; en résolvant le problème sur un réseau quatre fois plus lâche, on utilise de nouveau la duplication de la maille et l'augmentation de τ, et ainsi de suite. Dans ses expériences, Fédorenko utilise un réseau non pas deux mais trois fois plus lâche. Pour $M \approx 100$ deux augmentations de ce genre suffisent. Supposons pour plus de simplicité que $M = 2^k$, i.e. M est une puissance de deux.

Faisons plusieurs pas d'itérations (2) sur le réseau initial afin de « lisser » l'erreur. Celle-ci étant inconnue, on suit ce processus sur le résidu $\Delta_h u^p - \varphi$ qui est lissé lui aussi. Stockons le résultat $U = u^p$. Considérons pour la correction v le problème sur un réseau plus lâche, faisons plusieurs itérations (2*) dans le but de lisser « la correction de la correction » et stockons dans la mémoire le résultat $\tilde{V}^*$ (il occupe quatre fois moins de place que U). La correction de $\tilde{V}^*$ se calcule en considérant le problème sur un réseau deux fois moins serré, en effectuant plusieurs itérations de pas $\tau^{**} = 4\tau^* = 16\tau$ et en plaçant le résultat $\tilde{V}^{**}$ en mémoire. Ce processus de calcul des corrections sur des réseaux dont la maille continue d'augmenter de deux fois est répété k fois tant qu'on n'obtient pas le réseau le plus lâche et la correction $\tilde{V}^{(k*)}$.

Procédons ensuite en sens inverse. A partir du réseau le plus lâche interpolons la dernière correction $\tilde{V}^{(k*)}$ sur un réseau deux fois plus fin, portons cette correction interpolée dans $\tilde{V}^{((k-1)*)}$ et effectuons plusieurs itérations pour amortir l'erreur due à l'interpolation. Interpolons le résultat de ces itérations sur un réseau deux fois plus fin, précisons, ce résultat aidant, la correction $\tilde{V}^{((k-2)*)}$ conservée en mémoire pour ce réseau, faisons plusieurs itérations et interpolons. A l'avant-dernier pas on obtient, en corrigeant $\tilde{V}^*$ et en itérant plusieurs fois, la correction V^* qu'on interpole sur le réseau de départ. Plusieurs itérations (2) sur $U - V$ nous donnent enfin le résultat.

CINQUIÈME PARTIE

STABILITÉ DES PROBLÈMES AUX LIMITES AUX DIFFÉRENCES ÉVOLUTIFS EN TANT QUE BORNE DES NORMES DES PUISSANCES D'UN OPÉRATEUR

Dans les parties précédentes du livre nous avons accordé une grande attention à la stabilité des problèmes aux limites aux différences $L_h u^{(h)} = f^{(h)}$. Nous avons notamment étudié la stabilité de certains schémas aux différences approchant le problème de Dirichlet relatif à l'équation de Poisson. Ce problème est stationnaire : sa solution ne dépend pas du temps. Nous nous sommes cependant centrés sur des problèmes d'évolution pour les processus se déroulant dans le temps tels que la propagation de la chaleur ou des ondes. Les procédés d'étude des problèmes aux limites aux différences évolutifs sont mieux élaborés, ce qui s'explique en partie par la possibilité qui s'offre souvent de considérer des états stationnaires comme le résultat de la stabilisation des processus évoluant dans le temps.

En étudiant la stabilité des problèmes aux différences évolutifs nous avons utilisé le principe du maximum, les inégalités de l'énergie, les critères spectraux et d'autres procédés. Toutes ces techniques utilisent implicitement la structure particulière des schémas aux différences évolutifs en vertu de laquelle la solution $u^{(h)}$ est donnée à l'un ou à plusieurs niveaux de temps initiaux, puis elle se calcule pas à pas aux niveaux suivants. Ce caractère stratifié des schémas sera traduit par leur écriture : nous ferons correspondre au schéma aux différences un opérateur linéaire R_h qui permet de passer des valeurs connues à un temps donné à des valeurs inconnues au niveau suivant. Le choix de cet opérateur peut être divers. Nous allons le choisir de façon que la stabilité du schéma soit équivalente à la borne des normes de ses puissances. Cela permettra un point de vue unique sur les procédés examinés d'étude de la stabilité des problèmes aux limites aux différences évolutifs qui seront considérés comme ceux d'étude des propriétés des puissances de R_h. Moyennant ce choix de R_h, on formulera également la notion de spectre d'une famille d'opérateurs aux différences et le critère spectral de stabilité des problèmes aux limites aux différences non auto-adjoints.

CHAPITRE 12

CONSTRUCTION DE L'OPÉRATEUR DE PASSAGE

§38. Structure stratifiée des solutions des problèmes évolutifs

Dans tous les exemples ci-dessus de schémas aux différences évolutifs

$$L_h u^{(h)} = f^{(h)} \tag{1}$$

on a défini les valeurs de la solution $u^{(h)}$ à un ou à plusieurs niveaux de départ du réseau. Quant à ses valeurs aux niveaux suivants, elles se sont définies de proche en proche à partir des équations formant le problème aux limites (1). Nous entendons par *niveau* l'ensemble des points du réseau D_h situés sur la droite (ou dans le plan) $t = \text{const}$. Nous supposerons dans la suite que les schémas considérés (1) possèdent ce caractère stratifié.

EXEMPLE 1. Soit le schéma aux différences

$$\left.\begin{aligned} \frac{u_m^{p+1}-u_m^p}{\tau} &= \Lambda_{xx}u_m^p+\varphi(x_m, t_p), \\ u_0^{p+1} = \psi_1(t_{p+1}),\ u_M^{p+1} &= \psi_2(t_{p+1}),\ u_m^0 = \psi(x_m), \\ m = 0, 1, \ldots, M;\ p &= 0, 1, \ldots, [T/\tau]-1, \end{aligned}\right\} \tag{2}$$

approchant le problème de propagation de la chaleur

$$\left.\begin{aligned} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}+\varphi(x, t), &\quad 0 < x < 1,\ 0 < t < T, \\ u(0, t) = \psi_1(t), \quad u(1, t) = \psi_2(t), &\ u(x, 0) = \psi(x), \\ 0 < x < 1. & \end{aligned}\right\} \tag{3}$$

Connaissant les valeurs de $u^{(h)}$ aux points du niveau $t = t_p = p\tau$, i.e. connaissant la fonction discrète

$$u^p = \{u_m^p\}, \quad m = 0, 1, \ldots, M, \tag{4}$$

de l'argument m, on peut calculer successivement les valeurs des fonctions $u^{p+1} = \{u_m^{p+1}\}$, u^{p+2}, ... à l'aide de la formule

$$u_m^{p+1} = (1-2r)u_m^p+r(u_{m-1}^p+u_{m+1}^p)+\tau\varphi(x_m, t_p). \tag{5}$$

La fonction $u^0 = \{u_m^0\} = \{\psi(x_m)\}$ est donnée.

Ainsi, la solution $u^{(h)}$ définie sur le réseau à deux dimensions

$$(x_m, t_p) = (mh, p\tau), \quad m = 0, 1, \ldots, M;\ p = 0, 1, \ldots, [T/\tau] \tag{6}$$

du plan Oxt s'est décomposée de façon naturelle en une suite de fonctions

$$u^0, u^1, \ldots, u^p, \quad p = [T/\tau], \tag{7}$$

définies sur les réseaux unidimensionnels. Les réseaux à une dimension sur lesquels sont définis u^p, $p = 0, 1, \ldots, p$, sont les mêmes pour tous les p (fig. 46,a), de sorte qu'il s'agit en fait de divers exemplaires d'un même réseau (fig. 46,b).

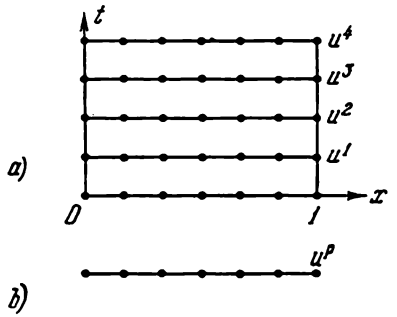


Fig. 46

Considérons l'espace vectoriel U' de fonctions définies sur ce dernier réseau (fig. 46,b). Il comprend en particulier les fonctions u^p, $p = 0, 1, \ldots, p$. Cet espace vectoriel étant supposé normé, la norme de l'élément $u = \{u_0, u_1, \ldots, u_M\}$ peut être définie par exemple par l'une des égalités

$$\left.\begin{aligned} \|u\| &= \max_m |u_m|, \\ \|u\| &= \left(h \sum_{m=0}^{M} |u_m|^2\right)^{1/2}. \end{aligned}\right\} \tag{8}$$

La définition de la stabilité et de la convergence fait appel à la norme $\|u^{(h)}\|_{U_h}$ de la solution du problème (1). Nous n'utiliserons que des normes $\|u^{(h)}\|_{U_h}$ qui tiennent compte du caractère stratifié de la solution $u^{(h)}$, à savoir

$$\|u^{(h)}\|_{U_h} = \max_p \|u^p\|, \tag{9}$$

p parcourant les valeurs $p = 0, 1, \ldots, [T/\tau]$, i.e. les valeurs telles que les domaines de définition des fonctions $u^p = \{u_m^p\}$ appartiennent à celui bidimensionnel de $u^{(h)}$.

Exemple 2. Considérons l'équation aux différences

$$\left.\begin{aligned} &\frac{u_m^{p+1} - 2u_m^p + u_m^{p-1}}{\tau^2} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \varphi(x_m, t_p), \\ &m = 0, \pm 1, \ldots; \qquad p = 1, 2, \ldots, [T/\tau]-1, \end{aligned}\right\} \tag{10}$$

l'analogue aux différences de l'équation différentielle

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = \varphi(x, t), \quad 0 < t < T, \ -\infty < x < \infty. \tag{11}$$

Contrairement à l'exemple 1, la solution $u^{(h)}$ de (10) n'est pas définie par ses valeurs aux points du niveau $t = p\tau$, et pour la définir il faut en connaître les valeurs à deux niveaux $t = p\tau$ et $t = (p+1)\tau$, i.e. la fonction vectorielle (fig. 47,a)

$$u^{(p)} = \begin{pmatrix} \ldots & u_{-1}^{p+1} & u_0^{p+1} & u_1^{p+1} & \ldots \\ \ldots & u_{-1}^p & u_0^p & u_1^p & \ldots \end{pmatrix}.$$

u^{p+1}, u^{p+2}, et ainsi de suite, se définissent à partir des valeurs de u^p en vertu

de l'équation (10). On prend donc pour espace U'_h l'espace de fonctions vectorielles (fig. 47,*b*)

$$u = \begin{pmatrix} \dots & b_{-1} & b_0 & b_1 & \dots \\ \dots & a_{-1} & a_0 & a_1 & \dots \end{pmatrix}$$

muni d'une norme $\|u\|$. Signalons quelques propriétés de celle-ci.

La solution de l'équation différentielle (11) est définie par deux fonctions

$$u(x, t_0) \quad \text{et} \quad \frac{\partial u(x, t)}{\partial t}$$

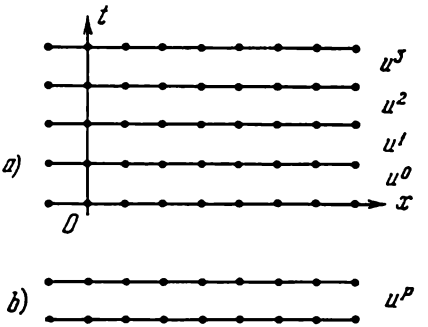


Fig. 47

dont les analogues aux différences sont les fonctions discrètes

$$\dots, u^p_{-1}, u^p_0, u^p_1, \dots$$

et

$$\dots, \frac{u^{p+1}_{-1} - u^p_{-1}}{\tau}, \quad \frac{u^{p+1}_0 - u^p_0}{\tau}, \quad \frac{u^{p+1}_1 - u^p_1}{\tau}, \dots$$

Toute norme naturelle dans l'espace U'_h doit donc dépendre des deux dernières fonctions. On pose par exemple

$$\|u\| = \sup_m |a_m| + \sup_m \left| \frac{b_m - a_m}{\tau} \right|$$

ou

$$\|u\| = \left[h \sum_m \left(|a_m|^2 + \frac{1}{\tau^2} |b_m - a_m|^2 \right) \right]^{1/2}.$$

Une fois l'espace U'_h muni d'une norme, on met automatiquement par la formule (9) une norme sur l'espace U_h de fonctions définies sur le réseau bidimensionnel :

$$\|u^{(h)}\|_{U_h} = \max_p \|u^p\|.$$

Le numéro p parcourt toutes les valeurs $p = 0, 1, \dots, [T/\tau]$ pour lesquelles les domaines de définition des fonctions vectorielles discrètes appartiennent au domaine bidimensionnel de définition de $u^{(h)}$.

Avec l'hypothèse faite d'utiliser toujours les normes de la forme (9), l'inégalité

$$\|u^{(h)}\|_{U_h} \leqslant c \|f^{(h)}\|_{F_h},$$

qui signifie qu'il y a stabilité si l'opérateur L_h est linéaire, équivaut à l'inégalité

$$\|u^p\| \leqslant c \|f^{(h)}\|_{F_h}$$

pour tous les p pour lesquels u^p est défini. Ce fait s'avère utile dans l'étude de la stabilité.

Exercices

1. Indiquer l'espace U_h' pour le schéma aux différences

$$\left.\begin{aligned} &\frac{u_{mn}^{p+1}-u_{mn}^{p}}{\tau} = \Lambda_{xx}u_{mn}^{p}+\Lambda_{yy}u_{mn}^{p}+\varphi(x_m, y_n, t_p),\\ &u_{mn}^{p}|_{\Gamma} = 0, \quad u_{mn}^{0} = \psi(x_m, y_n),\\ &m, n = 1, 2, \ldots, M-1;\ Mh = 1,\\ &p = 0, 1, \ldots, [T/\tau]-1, \end{aligned}\right\}$$

$(x_m, y_n, t_p) = (mh, nh, p\tau)$, Γ est formé des points du réseau qui sont situés sur la surface latérale du parallélépipède $0 \leqslant x, y \leqslant 1$, $0 \leqslant t \leqslant T$.

2. Indiquer l'espace U_h' pour le schéma de décomposition aux différences

$$\left.\begin{aligned} &\frac{u_{mn}^{p+1}-\tilde{u}_{mn}}{\tau} = \Lambda_{xx}u_{mn}^{p+1}+\varphi(x_m, y_n, t_p),\\ &\frac{\tilde{u}_{mn}-u_{mn}^{p}}{\tau} = \Lambda_{yy}\tilde{u}_{mn},\\ &u_{mn}^{p}|_{\Gamma} = \tilde{u}_{mn}|_{\Gamma} = 0, \quad u_{mn}^{0} = \psi(x_m, y_n),\\ &m, n = 1, 2, \ldots, M-1;\ Mh = 1;\ p = 0, 1, \ldots, [T/\tau]-1; \end{aligned}\right\}$$

Γ est la surface latérale du parallélépipède $0 \leqslant x, y \leqslant 1$, $0 \leqslant t \leqslant T$.

§39. Ecriture des problèmes aux limites aux différences sous forme $u^{p+1} = R_h u^p + \tau\varrho^p$

1. Forme canonique. Mettons le schéma aux différences

$$\left.\begin{aligned} &\frac{u_m^{p+1}-u_m^{p}}{\tau}-\frac{u_{m+1}^{p}-u_m^{p}}{h} = \varphi_m^p, \quad m = 0, \pm 1, \ldots,\\ &u_m^0 = \psi_m, \quad p = 0, 1, \ldots, [T/\tau]-1, \end{aligned}\right\}$$

relatif au problème

$$u_t-u_x = \varphi(x, t), \quad 0 < t < T,$$
$$u(x, 0) = \psi(x), \quad -\infty < x < \infty,$$

sous forme

$$\left.\begin{aligned} &u^{p+1} = \{(1-r)u_m^p+ru_{m+1}^p\}+\tau\varphi_m^p,\\ &u_m^0 = \psi_m, \quad r = \tau/h. \end{aligned}\right\} \tag{1}$$

En posant

$$v_m^{p+1} = (1-r)u_m^p+ru_{m+1}^p, \quad \varrho_m^p = \varphi_m^p, \tag{2}$$

récrivons-le comme suit :

$$u^{p+1} = v^{p+1}+\tau\varrho^p, \quad u_m^0 = \psi_m.$$

Le terme v^{p+1} est bien défini par $u^p = \{u_m^p\}$ et on peut donc l'écrire

$$v^{p+1} = R_h u^p,$$

R_h étant un opérateur qui fait correspondre à chaque fonction discrète $u^p \in U'_h$ la fonction discrète $v^{p+1} \in U'_h$ par la formule (2). L'écriture (1) devient alors

$$\left.\begin{aligned} u^{p+1} &= R_h u^p + \tau \varrho^p, \\ u^0 &\text{ donné.} \end{aligned}\right\} \tag{3}$$

Dans le présent paragraphe nous illustrerons sur d'autres exemples la réduction des problèmes d'évolution de la forme

$$L_h u^{(h)} = f^{(h)} \tag{4}$$

à la forme (3). Nous établirons que si cette réduction satisfait à certaines exigences naturelles, la stabilité de (4) sur le segment $0 \leqslant t \leqslant T$ équivaut aux inégalités

$$\| R_h^p \| < K, \quad p = 1, 2, \ldots, [T/\tau], \tag{5}$$

K étant une constante indépendante de h, et nous ramènerons donc l'étude de la stabilité à l'évaluation des quantités $\| R_h^p \|$, i.e. des normes des puissances de l'*opérateur de passage* R_h.

Des constructions et analyses analogues ont été faites aux paragraphes 15 et 16. Rappelons qu'au §38 l'étude de la stabilité a été ramenée à l'examen des inégalités

$$\| u^p \| \leqslant c \| f^{(h)} \|_{F_h}. \tag{6}$$

Nous avons notamment établi que la stabilité équivaut à l'existence d'un nombre c indépendant de h et $f^{(h)} \in F_h$ et tel que l'inégalité (6) soit vraie quels que soient p, $p = 1, 2, \ldots, [T/\tau]$.

Passons maintenant à l'exécution du plan esquissé et commençons par un exemple de réduction d'un schéma aux différences à la forme (3).

Soit le schéma

$$\left.\begin{aligned} &\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \varphi_m^p, \\ &p = 0, 1, \ldots, [T/\tau]-1; \; m = 1, 2, \ldots, M-1, \\ &u_0^p = \psi_1(t_p), \quad u_M^p = \psi_2(t_p), \quad p = 0, 1, \ldots, [T/\tau], \\ &u_m^0 = \psi(mh), \; m = 0, 1, \ldots, M. \end{aligned}\right\} \tag{7}$$

Il est clair que les conditions de compatibilité $\psi_1(0) = \psi(0)$, $\psi_2(0) = \psi(1)$ doivent être remplies. En vertu des conditions du problème $u^0 = \{u_m^0\}$ est donné et les fonctions $u^1, u^2, \ldots$ se calculent de proche en proche. A cette fin il faut récrire l'équation aux différences du schéma (7) :

$$u_m^{p+1} = (1-2r)u_m^p + r(u_{m-1}^p + u_{m+1}^p) + \tau\varphi_m^p,$$

$$r = \frac{\tau}{h^2}, \quad m = 1, 2, \ldots, M-1; \quad p = 0, 1, \ldots, [T/\tau]-1,$$

et utiliser les égalités

$$u_0^{p+1} = \psi_1(t_{p+1}), \quad u_M^{p+1} = \psi_2(t_{p+1}).$$

Prenons donc pour U'_h l'espace de fonctions discrètes

$$u = \{u_0, u_1, \ldots, u_M\}$$

avec pour norme

$$\|u\| = \max_m |u_m|.$$

Mettons le problème aux limites aux différences sous la forme

$$\left.\begin{aligned} &u^{p+1} = R_h u^p + \tau \varrho^p, \\ &u^0 \text{ donné}, \end{aligned}\right\} \tag{8}$$

R_h désignant un opérateur qui fait correspondre à tout élément $a = \{a_m\}$ de U'_h un élément $b = \{b_m\}$ du même espace conformément aux formules

$$\left.\begin{aligned} &b_0 = a_0, \\ &b_m = (1-2r)a_m + r(a_{m-1} + a_{m+1}), \quad m = 1, 2, \ldots, M-1, \\ &b_M = a_M. \end{aligned}\right\} \tag{9}$$

Pour ce choix de l'opérateur R_h la fonction discrète ϱ^p de U'_h

$$\varrho^p = \{\varrho_0^p, \varrho_1^p, \ldots, \varrho_M^p\}$$

se définit par la formule

$$\varrho^p = \left(\frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}, \varphi_1^p, \ldots, \varphi_{M-1}^p, \frac{\psi_2(t_{p+1}) - \psi_2(t_p)}{\tau}\right),$$
$$p = 0, 1, \ldots, [T/\tau]-1.$$

Nous avons ainsi réduit le schéma aux différences considéré (7) à la forme (3). Nous nous proposons d'utiliser cette écriture de problèmes aux limites aux différences pour étudier la stabilité. Mais pour que les inégalités (6) traduisant la stabilité aient un sens, il faut définir la norme $\|f^{(h)}\|_{F_h}$. Dans notre exemple le problème (7) s'écrit comme (4) à condition de poser

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - \dfrac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2}, & m = 1, 2, \ldots, M-1, \\ u_0^{p+1}, \quad p = 0, 1, \ldots, [T/\tau]-1, \\ u_M^{p+1}, \quad p = 0, 1, \ldots, [T/\tau]-1, \\ u_m^0, \quad m = 0, 1, \ldots, M, \end{cases}$$

$$f^{(h)} \equiv \begin{cases} \varphi(x_m, t_p), & m = 1, 2, \ldots, M-1, \\ \psi_1(t_{p+1}), & p = 0, 1, \ldots, [T/\tau]-1, \\ \psi_2(t_{p+1}), & p = 0, 1, \ldots, [T/\tau]-1, \\ \psi(x_m), & m = 0, 1, \ldots, M. \end{cases}$$

La norme $\|f^{(h)}\|_{F_h}$ sera définie par l'égalité

$$\|f^{(h)}\|_{F_h} = \max_{m,p} |\varphi(x_m, t_p)| + \max_m |\psi(x_m)| + \max_p \left| \frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau} \right| + \\ + \max_p \left| \frac{\psi_2(t_{p+1}) - \psi_2(t_p)}{\tau} \right|. \quad (10)$$

2. Stabilité en tant que borne uniforme des normes des puissances de l'opérateur R_h. Formulons deux conditions qu'il faut imposer à la réduction d'un schéma aux différences (4) à la forme (3) pour pouvoir affirmer que si les inégalités (5) sont vraies, il y a stabilité.

CONDITION 1° *On a les inégalités*

$$\|\varrho^p\| \leqslant K_1 \|f^{(h)}\|_{F_h},$$

avec K_1 indépendant de h et $f^{(h)}$ et p parcourant toutes les valeurs pour lesquelles ϱ^p a un sens.

CONDITION 2° *On a l'estimation*

$$\|u^0\| \leqslant K_2 \|f^{(h)}\|_{F_h},$$

avec K_2 une constante indépendante de h et $f^{(h)}$.

Les conditions 1° et 2° exigent un choix concordant des normes dans les espaces U_h' et F_h et de R_h (la forme du vecteur ϱ^p dépend de façon unique du choix de l'opérateur R_h). Notons que nous avons bien ces conditions dans l'exemple considéré de réduction du schéma aux différences (7) à la forme (3). Pour s'en convaincre il suffit de comparer les normes des fonctions discrètes ϱ^p et u^0

$$\|\varrho^p\| = \max_m |\varrho_m^p| = \max \left(\left| \frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau} \right|, \quad |\varphi(x_1, t_p)|, \ldots \right. \\ \left. \ldots, |\varphi(x_{M-1}, t_p)|, \left| \frac{\psi_2(t_{p+1}) - \psi_2(t_p)}{\tau} \right| \right),$$

$$\|u^0\| = \max_m |u_m^0| = \max_m |\psi(x_m)|$$

avec l'égalité (10) définissant la norme $\|f^{(h)}\|_{F_h}$.

Les nombres K_1 et K_2 de notre exemple sont supposés égaux à l'unité.

Démontrons que si la réduction du problème aux limites aux différences (4) à la forme (3) remplit les conditions 1° et 2°, la condition suffisante de stabilité du schéma (4) est qu'on ait les estimations (5). Il faut montrer que les estimations

$$\|u^p\| \leqslant K_3 \|f^{(h)}\|_{F_h},$$

K_3 étant un nombre indépendant de h et $f^{(h)}$, sont satisfaites pour tous les p, $p = 0, 1, \ldots, p_0$, pour lesquels les domaines de définition des fonctions u^p appartiennent à celui de la solution $u^{(h)}$.

L'équation

$$u^{p+1} = R_h u^p + \tau \varrho^p, \quad p = 0, 1, \ldots, p_0 - 1,$$

entraîne

$$u^p = R_h u^{p-1} + \tau \varrho^{p-1} = R_h(R_h u^{p-2} + \tau \varrho^{p-2}) + \tau \varrho^{p-1} =$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$= R_h^p u^0 + \tau(R_h^{p-1}\varrho^0 + R_h^{p-2}\varrho^1 + \ldots + \varrho^{p-1}), \quad p = 0, 1, \ldots, p_0. \quad (11)$$

Par hypothèse,

$$\|R_h^l\| \leqslant K, \quad l = 0, 1, \ldots, [T/\tau].$$

Cette inégalité et l'égalité (11) impliquent l'estimation

$$\|u^p\| \leqslant K\|u^0\| + \tau K(\|\varrho^0\| + \|\varrho^1\| + \ldots + \|\varrho^{p-1}\|). \quad (12)$$

Compte tenu des conditions 1° et 2° en vertu desquelles

$$\|\varrho^l\| \leqslant K_1 \|f^{(h)}\|_{F_h}, \quad l = 0, 1, \ldots, p_0 - 1,$$

$$\|u^0\| \leqslant K_2 \|f^{(h)}\|_{F_h},$$

l'estimation (12) peut être remplacée par

$$\|u^p\| \leqslant (KK_2 + \tau pKK_1)\|f^{(h)}\|_{F_h} \leqslant K(K_2 + TK_1)\|f^{(h)}\|_{F_h} = c\,\|f^{(h)}\|_{F_h},$$

où le nombre $c = K(K_2 + TK_1)$ est indépendant de h et $f^{(h)}$.

La stabilité est démontrée.

Utilisons le critère suffisant de stabilité que nous avons démontré et montrons la stabilité du schéma aux différences (7) pour $r = \tau/h^2 \leqslant 1/2$.

Assurons-nous notamment que l'opérateur R_h que nous avons introduit par les formules (9), lors de la réduction du schéma aux différences (7) à la forme canonique (3), vérifie l'inégalité $\|R_h\| \leqslant 1$ et donc l'inégalité $\|R_h^p\| \leqslant \|R_h\|^p \leqslant 1$. Pour $r \leqslant 1/2$ on a les estimations

$$|b_0| = |a_0| \leqslant \max_m |a_m| = \|a\|,$$

$$|b_m| = |(1-2r)a_m + r(a_{m-1} + a_{m+1})| \leqslant (1 - 2r + 2r) \max_m |a_m| = \|a\|,$$

$$|b_M| = |a_M| \leqslant \max_m |a_m| = \|a\|,$$

d'où il découle que $\|b\| = \|R_h a\| \leqslant \|a\|$, i.e. $\|R_h\| \leqslant 1$. Ainsi, le critère suffisant de stabilité est satisfait pour $r \leqslant 1/2$. On montre qu'il n'en est pas ainsi pour $r = \tau/h^2 > 1/2$. Une question se pose alors : la stabilité ne disparaît-elle également pas dans le cas général où les inégalités $\|R_h^p\| < K$, $p = 1, 2, \ldots, [T/\tau]$, cessent d'être justes ? Il s'avère qu'en effet les inégalités $\|R_h^p\| < K$ ne sont une condition nécessaire de stabilité que sous une condition 3° que nous allons énoncer sous une forme générale et qui est remplie dans l'exemple considéré.

CONDITION 3° Supposons que le problème aux limites aux différences (4) est ramené à la forme (3). Prenons une fonction $\bar{u}^0 \in U_h'$ et construisons les

fonctions discrètes $\bar{u}^1, \bar{u}^2, \ldots, \bar{u}^p, \ldots$ d'après les formules de récurrence $\bar{u}^{p+1} = R_h\bar{u}^p$. La famille de fonctions discrètes $\{\bar{u}^p\}, p = 0, 1, \ldots, [T/\tau]$, dont chacune est un élément de U'_h, constitue une fonction discrète $\bar{u}^{(h)}$ de l'espace U_h. Calculons pour celle-ci $\bar{f}^{(h)}$,

$$\bar{f}^{(h)} \equiv L_h\bar{u}^{(h)}.$$

Nous dirons que *la réduction du schéma aux différences* (4) *à la forme canonique* (3) *satisfait à la condition* 3° *si l'on a une estimation de la forme*

$$\|\bar{f}^{(h)}\|_{F_h} \leqslant K_3\|\bar{u}^0\|,$$

où la constante K_3 est indépendante de $\bar{u}^0 \in U'_0$ et de h.

Convainquons-nous de ce que la réduction de (7) à la forme canonique (3) remplit la condition 3°. En effet, en nous donnant une fonction arbitraire $\bar{u}^0 = \{\bar{u}^0_m\}$, nous trouvons

$$\bar{\varphi}^p_m \equiv 0, \quad \bar{\psi}^p_1 = u^0_0, \quad \bar{\psi}^p_2 = u^0_M, \quad \bar{\psi}_m \equiv \bar{u}^0_m.$$

Pour notre choix des normes, nous avons l'égalité

$$\|\bar{f}^{(h)}\|_{F_h} = \|\bar{u}^0\|.$$

Démontrons maintenant que si, en ramenant le schéma aux différences (4) à la forme canonique (3), nous remplissons la condition 3°, pour que ce schéma soit stable sur le segment $0 < t < T$, il faut avoir les estimations (5) :

$$\|R^p_h\| < K, \quad p = 1, 2, \ldots, [T/\tau],$$

K étant un nombre indépendant de h.

Si le critère formulé n'est pas vérifié, il existe pour tout K des h, p_0 et une fonction $\bar{u}^0$ tels que

$$\|R^{p_0}_h\bar{u}^0\| > K\|u^0\| \quad (\|u^{p_0}\| > K\|u^0\|).$$

En construisant à partir de $\bar{u}^0$ les vecteurs $\bar{u}^p$ ($\|u^{p_0}\| > K\|u^0\|$) et en formant avec $\bar{u}^p$ la fonction discrète $\bar{u}^{(h)}$, nous concluons, compte tenu de la condition 3°, que pour cette fonction

$$\|\bar{f}^{(h)}\| \leqslant K_3\|\bar{u}^0\| \quad (\bar{f}^{(h)} = L_h\bar{u}^{(h)}).$$

De plus,

$$\|\bar{u}^{(h)}\|_{U_h} = \max_p \|\bar{u}^p\| \geqslant \|\bar{u}^{p_0}\| > K\|\bar{u}^0\|.$$

Il est donc clair que

$$\|\bar{u}^{(h)}\|_{U_h} > \frac{K}{K_3}\|\bar{f}^{(h)}\|_{F_h}.$$

K étant arbitraire, cette inégalité signifie la stabilité.

Dressons le bilan du présent paragraphe. Nous avons montré qu'en

ramenant le schéma aux différences $L_h u^{(h)} = f^{(h)}$ à la forme (3)

$$\left.\begin{aligned} u^{p+1} &= R_h u^p + \tau \varrho^p \\ u^0 &\text{ donné,} \end{aligned}\right\}$$

on peut ensuite utiliser l'opérateur R_h pour étudier la stabilité. Nous avons notamment démontré le

THÉORÈME. *Si en ramenant le schéma aux différences* (4) *à la forme* (3) *on remplit la condition* 3°, *la condition nécessaire de stabilité est qu'on ait l'inégalité*

$$\| R_h^p \| < K, \quad p = 1, 2, \ldots, [T/\tau], \tag{13}$$

avec K un nombre indépendant de h. Si la réduction en question se fait dans les conditions 1° *et* 2°, *les estimations* (13) *suffisent pour avoir la stabilité.*

Nous voulons attirer l'attention du lecteur sur le fait qu'en général la décomposition de la fonction $u^{(h)}$ et la réduction du schéma aux différences à la forme (3) peuvent s'effectuer de diverses façons. Nous ne nous y arrêterons cependant pas (voir §14 où la discussion correspondante porte sur des schémas aux différences pour les équations différentielles ordinaires).

3. Exemple. Nous voulons terminer ce paragraphe en examinant le schéma aux différences implicite suivant :

$$\left.\begin{gathered} \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m-1}^{p+1} - 2u_m^{p+1} + u_{m+1}^{p+1}}{h^2} = \varphi_m^p, \\ m = 1, 2, \ldots, M-1; \quad p = 0, 1, \ldots, [T/\tau]-1, \\ u_m^0 = \psi_m, \quad m = 0, \ldots, M, \\ u_0^{p+1} = \psi_1(t_{p+1}), \quad u_M^{p+1} = \psi_2(t_{p+1}), \quad p = 0, 1, \ldots, [T/\tau]-1, \end{gathered}\right\} \tag{14}$$

pour le problème de la chaleur

$$\left.\begin{gathered} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = \varphi(x, t), \\ u(x, 0) = \psi(x), \\ u(0, t) = \psi_1(t), \quad u(1, t) = \psi_2(t), \\ 0 < t < T, \quad 0 < x < 1. \end{gathered}\right\} \tag{15}$$

Ce schéma a été étudié en détail au §28.

Acceptons comme u^p le vecteur $u^p = (u_0^p, u_1^p, \ldots, u_M^p)$ de norme $\| u^p \| = \max_m |u_m^p|$. La solution au $(p+1)$-ième niveau s'écrira sous forme de somme

$$u^{p+1} = v^{p+1} + \tau \varrho^p,$$

où

$$v^{p+1} = (v_0^{p+1}, v_1^{p+1}, \ldots, v_M^{p+1}) \quad \text{et} \quad \varrho^p = (\varrho_0^p, \varrho_1^p, \ldots, \varrho_M^p)$$

sont également solutions des systèmes auxiliaires d'équations

$$\left.\begin{aligned} v_0^{p+1} &= u_0^p = \psi_1(t_p), \\ rv_{m+1}^{p+1}-(1+2r)v_m^{p+1}+rv_{m-1}^{p+1} &= -u_m^p, \\ m &= 1, 2, \ldots, M-1, \\ v_M^{p+1} &= u_M^p = \psi_2(t_p), \end{aligned}\right\} \tag{16}$$

$$\left.\begin{aligned} \varrho_0^p &= \frac{\psi_1(t_{p+1})-\psi_1(t_p)}{\tau}, \\ r\varrho_{m+1}^p-(1+2r)\varrho_m^p+r\varrho_{m-1}^p &= \varphi_m^p, \quad m = 1, 2, \ldots, M-1, \\ \varrho_M^p &= \frac{\psi_2(t_{p+1})-\psi_2(t_p)}{\tau}. \end{aligned}\right\} \tag{17}$$

Prenant le premier système pour définition de l'opérateur R_h posons

$$v^{p+1} = R_h u^p.$$

Si $\|f^{(h)}\|_{F_h}$ se définit toujours par l'égalité (10), la condition 1°

$$\|\varrho^p\| \leqslant K_1 \|f^{(h)}\|_{F_h}$$

découle de l'estimation

$$\|\varrho^p\| \leqslant \max\left[\left|\frac{\psi_1(t_{p+1})-\psi_1(t_p)}{\tau}\right|, \left|\frac{\psi_2(t_{p+1})-\psi_2(t_p)}{\tau}\right|, \max_m |\varphi_m^p|\right]$$

vérifiée par la solution $\{\varrho_m^p\}$ du système (17) en vertu de (7) du §4. Cela étant, on a $K_1 = 1$.

La condition 2°

$$\|u^0\| \leqslant K_2 \|f^{(h)}\|_{F_h}$$

est remplie elle aussi en vertu de (10) parce que $u_m^0 = \psi_m$, et il est évidemment possible de poser $K_2 = 1$. Au §28 nous avons démontré l'estimation

$$|v_m^{p+1}| \leqslant \max_m |u_m^p|$$

qu'on peut interpréter comme les inégalités

$$\|R_h u\| \leqslant \|u\|, \quad \|R_h\| \leqslant 1,$$

d'où

$$\|R_h^p\| \leqslant K = 1.$$

Nous avons reproduit le schéma de démonstration de la stabilité du §28 et montré qu'il coïncide avec le schéma proposé. L'exemple considéré est intéressant sous un autre rapport encore : il utilise un procédé assez ardu de construction du vecteur ϱ^p.

Exercices

1. Pour le système d'équations de l'acoustique

$$\frac{\partial \mathbf{u}}{\partial t}+A\frac{\partial \mathbf{u}}{\partial x}=\boldsymbol{\varphi}(x,t),\qquad -\infty<x<\infty,\ 0<t<T,$$
$$\mathbf{u}(x,0)=\boldsymbol{\phi}(x),\qquad -\infty<x<\infty,$$
$$A=\begin{pmatrix}0&1\\1&0\end{pmatrix},\qquad \mathbf{u}(x,t)=\begin{pmatrix}v(x,t)\\w(x,t)\end{pmatrix},$$
$$\boldsymbol{\varphi}(x,t)=\begin{pmatrix}\varphi_1(x,t)\\\varphi_2(x,t)\end{pmatrix},\qquad \boldsymbol{\phi}(x)=\begin{pmatrix}\psi_1(x)\\\psi_2(x)\end{pmatrix}$$

réduire à la forme canonique (3) le schéma

$$\left.\begin{aligned}&\frac{\mathbf{u}_m^{p+1}-\mathbf{u}_m^p}{\tau}+A\frac{\mathbf{u}_{m+1}^p-\mathbf{u}_{m-1}^p}{2h}-\\&-\frac{\tau}{2h^2}A^2(\mathbf{u}_{m+1}^p-2\mathbf{u}_m^p+\mathbf{u}_{m-1}^p)=\boldsymbol{\varphi}(x_m,t_p),\\&\mathbf{u}_m^0=\boldsymbol{\phi}(x_m)\end{aligned}\right\}\qquad(*)$$

en faisant $\mathbf{u}^p=\{\mathbf{u}_m^p\}$. Vérifier qu'en choisissant les normes par les formules

$$\|\mathbf{u}^{(h)}\|_{U_h}=\max_p\|\mathbf{u}^p\|,\qquad \|\mathbf{f}^{(h)}\|_{F_h}=\max\left[\|\boldsymbol{\phi}\|,\ \max_p\|\boldsymbol{\varphi}^{(p)}\|\right],$$

où

$$\|\mathbf{u}^p\|^2=\sum_m(|v_m^p|^2+|w_m^p|^2),\qquad \|\boldsymbol{\phi}\|^2=\sum_m(|\psi_1(mh)|^2+|\psi_2(mh)|^2),$$
$$\|\boldsymbol{\varphi}^{(p)}\|^2=\sum_m(|\varphi_1(x_m,t_p)|^2+|\varphi_2(x_m,t_p)|^2),$$

nous vérifions les conditions de réduction 1° à 3°.

Démontrer que lorsque $r\leqslant 1$, le schéma aux différences (*) est stable et qu'il est instable pour $r>1$.

INDICATION. Pour évaluer les normes $\|R_h^p\|$ passer en variables

$$I_m^{(1)}=v_m+w_m,\qquad I_m^{(2)}=v_m-w_m,$$

dites invariants de Riemann, et utiliser le critère spectral du n° 4 §25.

2. En posant

$$u^p=\begin{Bmatrix}v_m^{p+1}\\v_m^p\end{Bmatrix},$$
$$\|u^p\|^2=\sum_m|v_m^p|^2+\sum_m\left|\frac{v_m^{p+1}-v_m^p}{\tau}\right|^2,$$
$$\|\varphi^p\|^2=\sum_m|\varphi_m^p|^2,\qquad \|\psi_k\|^2=\sum|\psi_k(x_m)|^2,$$
$$\|u^{(h)}\|_{U_h}=\max_p\|u^p\|,$$
$$\|f^{(h)}\|_{F_h}=\|\psi_0\|+\|\psi_1\|+\max_p\|\varphi^p\|$$

ramener à la forme canonique (3) le schéma aux différences

$$\left.\begin{aligned}&\frac{v_m^{p+1}-2v_m^p+v_m^{p-1}}{\tau^2}-\frac{v_{m+1}^p-2v_m^p+v_{m-1}^p}{h^2}=\varphi_m^p,\\&p=1,\ldots,[T/\tau]-1,\\&v_m^0=\psi_0(x_m),\qquad v_m^1=\tau\psi_1(x_m)+\varphi_0(x_m),\qquad m=0,\pm1,\ldots,\end{aligned}\right\}$$

qui approche le problème de Cauchy

$$\frac{\partial^2 v}{\partial t^2} - \frac{\partial^2 v}{\partial x^2} = \varphi(x, t), \quad -\infty < x < \infty, \quad 0 < t < T,$$

$$v(x, 0) = \psi_0(x), \quad \frac{\partial v(x, 0)}{\partial t} = \psi_1(x), \quad -\infty < x < \infty.$$

a) Vérifier que les conditions 1° à 3° sont remplies.

b) Démontrer la stabilité du schéma pour $\tau/h = r \leqslant 1$ et son instabilité pour $\tau/h = r > 1$.

§40. Utilisation des solutions particulières pour construire l'opérateur de passage

Nous avons parlé au §39 de la réduction du problème aux limites aux différences

$$L_h u^{(h)} = f^{(h)} \tag{1}$$

à la forme

$$\left.\begin{aligned} u^{p+1} &= R_h u^p + \tau \varrho^p, \\ u^0 &\text{ donné.} \end{aligned}\right\} \tag{2}$$

Ceci étant, l'opérateur R_h est choisi de différentes manières. En réduisant à la forme (2) on veut juger de la stabilité d'après les estimations des $\|R_h^p\|$. On a montré que l'estimation

$$\|R_h^p\| < K, \quad p = 1, 2, \ldots, [T/\tau], \tag{3}$$

garantit la stabilité si R_h et les normes sont choisis de façon qu'on ait les conditions

1° $$\|\varrho^p\| \leqslant K_1 \|f^{(h)}\|_{F_h},$$

où p parcourt toutes les valeurs pour lesquelles ϱ^p est défini ;

2° $$\|u^0\| \leqslant K_2 \|f^{(h)}\|_{F_h}.$$

Dans les exemples du §39 on a réussi à choisir un R_h suffisamment simple tout en garantissant les conditions 1° et 2°. Il peut cependant arriver (un exemple en sera examiné au présent paragraphe) que la condition 1° soit trop sévère, ce qui exclut le choix d'un R_h suffisamment simple.

Nous allons montrer que l'estimation (3) reste une condition suffisante de stabilité si la condition 1° est remplacée par une contrainte 1* moins rigide, ce qui permet, en réduisant le schéma aux différences à la forme (2), de choisir un R_h d'autant plus simple que l'information sur les solutions du problème proposé (1) est plus riche. En particulier, la structure d'un R_h tel qu'il rende possible la réduction à la forme (2) se simplifie si on connaît certaines solutions particulières de l'équation aux différences figurant dans le problème aux limites (1). La démonstration de l'inégalité (3) entraînant la stabilité se simplifie en conséquence.

Plus loin nous l'illustrerons par des exemples. Pour le moment formulons la condition 1*.

Soit $z^{(h)}$ une fonction discrète de U_h qui dépend en général de $f^{(h)}$. En ramenant le schéma aux différences (1) à la forme canonique (2) nous décomposons la fonction $u^{(h)}$ en des fonctions $u^p \in U'_h$. Procédons de même en ce qui concerne $z^{(h)}$, décomposons-la en $z^p \in U'_h$ et supposons que z^p vérifient des inégalités de la forme

$$\|z^p\| \leqslant \tilde{K}\|f^{(h)}\|_{F_h}, \tag{4}$$

$\tilde{K}$ étant une constante et p prenant les valeurs $p = 0, 1, \ldots, p_0$ pour lesquelles z^p est défini.

CONDITION 1*. *Il existe une fonction $z^{(h)}$ vérifiant les inégalités* (4) *et telle qu'on ait les estimations*

$$\left\| \varrho^p - \frac{1}{\tau}(z^{p+1} - R_h z^p) \right\| \leqslant K_1 \|f^{(h)}\|_{F_h}.$$

En prenant pour $z^{(h)}$ sa valeur indentiquement nulle, on remplit, en plus de 1*, la condition plus rigide 1°.

THÉORÈME. *Si le problème aux différences* (1) *s'écrit sous forme canonique* (2) *dans les conditions* 1* *et* 2°, *l'estimation* (3) *implique l'inégalité*

$$\|u^p\| \leqslant c\|f^{(h)}\|_{F_h} \tag{5}$$

signifiant qu'il y a stabilité. La constante c peut être choisie par la formule

$$c = K(2K_2 + 2\tilde{K} + TK_1) + \tilde{K}.$$

DÉMONSTRATION. Introduisons une fonction $w^{(h)} = u^{(h)} - z^{(h)}$ telle que l'équation

$$u^{p+1} = R_h u^p + \tau\varrho^p$$

entraîne l'égalité

$$w^{p+1} = R_h w^p + \tau\tilde{\varrho}^p, \tag{6}$$

où

$$\tilde{\varrho}^p = \varrho^p - \frac{1}{\tau}(z^{p+1} - R_h z^p).$$

En vertu de la condition 1* nous avons

$$\|\tilde{\varrho}^p\| \leqslant K_1 \|f^{(h)}\|_{F_h}.$$

Utilisons l'égalité (6) et l'inégalité (3) et trouvons sans peine (comme nous l'avons fait plus d'une fois)

$$\|w^p\| \leqslant K\|w^0\| + TK \max_p \|\tilde{\varrho}^p\| \leqslant K\|w^0\| + TKK_1\|f^{(h)}\|_{F_h}. \tag{7}$$

Ensuite,

$$\|w^0\| \leqslant 2(K_2 + \tilde{K})\|f^{(h)}\|_{F_h}. \tag{8}$$

Cela résulte des inégalités

$$\begin{aligned} \|w^0\| &= \|u^0 - z^0\| \leqslant \|u^0\| + \|z^0\|, \\ \|u^0\| &\leqslant K_2 \|f^{(h)}\|_{F_h}, \\ \|z^0\| &\leqslant \tilde{K} \|f^{(h)}\|_{F_h}, \end{aligned}$$

dont la deuxième coïncide avec la condition 2° et la troisième avec l'inégalité (4) lorsque $p = 0$.

La substitution de l'estimation (8) dans (7) donne

$$\|w^p\| \leqslant [2(K_2+\tilde{K})K+TKK_1] \|f^{(h)}\|_{F_h}.$$

Il reste à remarquer que

$$\begin{aligned} \|u^p\| &= \|w^p+z^p\| \leqslant \|w^p\|+\|z^p\| \leqslant \\ &\leqslant \{[2K(K_2+\tilde{K})+TKK_1]+\tilde{K}\} \|f^{(h)}\|_{F_h} \leqslant \\ &\leqslant [K(2K_2+2\tilde{K}+TK_1)+\tilde{K}] \|f^{(h)}\|_{F_h} = c \|f^{(h)}\|_{F_h}. \end{aligned}$$

La substitution de la condition 1* à 1° nous permet de diviser les difficultés que présente l'étude de la stabilité entre la construction d'un opérateur R_h tel que les normes de ses puissances s'évaluent sans trop de peine et la démonstration de l'existence de la fonction $z^{(h)}$. En exigeant dès le début que la condition 1* soit remplie pour $z^{(h)} = 0$, i.e. la condition 1°, nous imposons des contraintes rigides au choix de R_h. Il se peut que tout R_h choisi dans la condition 1° soit d'une forme compliquée, de sorte que l'évaluation des normes de ses puissances s'avère une opération trop ardue. D'autre part, le fait de choisir un R_h simple à l'extrême, disons unité, sans tenir compte du problème aux différences, laisse toutes les difficultés d'un seul côté (la vérification de la condition 1*, i.e. les évaluations de la fonction $z^{(h)}$ qu'il est alors le plus naturel de choisir telle qu'elle coïncide avec $u^{(h)}$). L'introduction d'un tel opérateur R_h et d'une telle fonction $z^{(h)}$ ne nous avancerait pas dans l'étude de la stabilité.

On prendra pour R_h l'opérateur le plus simple possible. R_h doit cependant tenir compte des propriétés du problème aux différences $L_h u^{(h)} = f^{(h)}$ et ce au point que la condition 1*, i.e. l'existence de la fonction $z^{(h)}$, soit suffisamment évidente. On arrive souvent à utiliser l'arbitraire dans le choix de R_h, arbitraire dû au remplacement de 1° par la condition moins restrictive 1*, pour faciliter la démonstration de la stabilité. On prend alors en qualité de $z^{(h)}$ des fonctions construites à partir des solutions des problèmes aux différences pour un second membre $f^{(h)}$ d'une forme spéciale.

Nous allons illustrer le procédé proposé par plusieurs exemples.

Exemple 1. Considérons le problème aux limites aux différences (1) de la forme

$$\left.\begin{aligned} \frac{u_m^{p+1}-u_m^p}{\tau} - \frac{u_{m+1}^p-u_m^p}{h} &= \varphi_m^p, \quad m = 0, 1, \ldots, M-1; \quad Mh = 1, \\ u_m^0 &= \psi(x_m), \quad m = 0, 1, \ldots, M, \\ u_M^p &= \psi_1(t_p), \quad p = 1, 2, \ldots, [T/\tau]. \end{aligned}\right\} \quad (9)$$

Ce problème approche le problème

$$\left.\begin{aligned} u_t - u_x &= \varphi(x, t) \quad 0 < x < 1, \quad 0 < t < T, \\ u(x, 0) &= \psi(x), \quad 0 < x < 1, \\ u(1, t) &= \psi_1(t), \quad 0 < t < T, \end{aligned}\right\} \tag{10}$$

pour les normes définies par

$$\|u^{(h)}\|_{U_h} = \max_p \max_m |u_m^p|,$$

$$\|f^{(h)}\|_{F_h} = \max_p \max_m |\varphi_m^p| + \max_m |\psi(x_m)| +$$

$$+ \max_p |\psi_1(t_p)| + \max_p \left| \frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau} \right|.$$

Pour réduire (9) à la forme canonique (2) posons

$$u^p = (u_0^p, u_1^p, \ldots, u_M^p), \qquad \|u\| = \max_m |u_m|.$$

L'opérateur R_h, $b = R_h a$, transformant l'élément $a = (a_0, a_1, \ldots, a_M)$ de l'espace U_h' en l'élément $b = (b_0, b_1, \ldots, b_M)$ du même espace, se définit par les égalités

$$\left.\begin{aligned} b_m &= (1-r)a_m + r a_{m+1}, \quad m = 0, 1, \ldots, M-1, \\ b_M &= 0, \qquad\qquad\qquad\quad r = \tau/h. \end{aligned}\right\}$$

On a alors évidemment

$$\varrho^p = \left(\varphi_0^p, \varphi_1^p, \ldots, \varphi_{M-1}^p, \frac{\psi_1(t_{p+1})}{\tau}\right).$$

Il est clair que la condition 1°

$$\|\varrho^p\| \leqslant K_1 \|f^{(h)}\|_{F_h}$$

n'est pas remplie parce que la dernière composante du vecteur ϱ^p est $\frac{\psi_1(t_{p+1})}{\tau}$ et croît pour $\tau \to 0$. (Il serait facile dans notre problème de définir un opérateur R_h tel que la condition 1° soit satisfaite. Il suffit pour cela de remplacer $b_M = 0$ dans la définition de R_h par $b_M = a_M$.) Il est pourtant facile de trouver un z^p tel que la condition 1*

$$\|\varrho^p - \frac{1}{\tau}(z^{p+1} - R_h z^p)\| \leqslant \tilde{K} \|f^{(h)}\|_{F_h}$$

se trouve vérifiée. Le premier membre de cette inégalité s'écrit comme

$$\left\| \left(\varphi_0^p, \varphi_1^p, \ldots, \varphi_{M-1}^p, \frac{\psi_1(t_{p+1})}{\tau}\right) - \frac{1}{\tau}(z^{p+1} - R_h z^p) \right\|.$$

Pour démontrer la condition 1* il suffit donc de construire une fonction $z^{(h)}$ vérifiant l'équation

$$\frac{1}{\tau}(z^{p+1} - R_h z^p) = \left(0, 0, \ldots, 0, \frac{\psi_1(t_{p+1})}{\tau}\right)$$

qui se récrit comme

$$z^{p+1} = R_h z^p + \tau\left(0, 0, \ldots, 0, \frac{\psi_1(t_{p+1})}{\tau}\right)$$

ou encore

$$\left.\begin{aligned} \frac{z_m^{p+1}-z_m^p}{\tau} - \frac{z_{m+1}^p - z_m^p}{h} &= 0, \quad m = 0, 1, \ldots, M-1, \\ z_M^{p+1} &= \psi_1(t_{p+1}). \end{aligned}\right\} \tag{11}$$

Si $\psi_1(t_{p+1})$ ne dépend pas de t, ce problème admet une solution stationnaire (indépendante de p)

$$z_m^p \underset{m}{\equiv} \psi_1 = \text{const}.$$

Dans le cas général $\psi_1^p = \psi_1(t_p)$ dépend de p, mais pour la norme $\|f^{(h)}\|_{F_h}$ bornée contenant le terme $\left|\frac{\psi_1(t_{p+1})-\psi_1(t_p)}{\tau}\right|$, il ne peut varier brusquement. Aussi, bien que la fonction $z^{(h)}$ définie par l'égalité

$$z_m^p \underset{m}{\equiv} \psi_1(t_p)$$

ne soit pas solution stationnaire (ce n'est en général pas une solution) du problème (11), elle satisfait « presque » aux relations (11). A savoir

$$\frac{1}{\tau}(z^{p+1} - R_h z^p) = \left(\frac{\psi_1(t_{p+1})-\psi_1(t_p)}{\tau}, \frac{\psi_1(t_{p+1})-\psi_1(t_p)}{\tau}, \ldots\right.$$
$$\left.\ldots, \frac{\psi_1(t_{p+1})-\psi_1(t_p)}{\tau}, \frac{\psi_1(t_{p+1})}{\tau}\right).$$

C'est pourquoi

$$\left\|\varrho^p - \frac{1}{\tau}(z^{p+1} - R_h z^p)\right\| = \left\|\left(\varphi_0^p - \frac{\psi_1(t_{p+1})-\psi_1(t_p)}{\tau},\right.\right.$$
$$\left.\left.\varphi_1^p - \frac{\psi_1(t_{p+1})-\psi_1(t_p)}{\tau}, \ldots, \varphi_{M-1}^p - \frac{\psi_1(t_{p+1})-\psi_1(t_p)}{\tau}, 0\right)\right\| \leqslant$$
$$\leqslant \|(\varphi_0^p, \varphi_1^p, \ldots, \varphi_{M-1}^p, 0)\| + \left|\frac{\psi_1(t_{p+1})-\psi_1(t_p)}{\tau}\right| \|(1, 1, \ldots, 1, 0)\| \leqslant \|f^{(h)}\|_{F_h}.$$

La condition 1° est remplie; $\bar{K} = 1$ et $z^{(h)}$ vérifient l'inégalité

$$\|z^p\| = \max_m |z_m^p| \leqslant \|f^{(h)}\|_{F_h}.$$

La condition 2°

$$\|u^0\| \leqslant K_2 \|f^{(h)}\|_{F_h}$$

l'est également :

$$\|u^0\| = \max_m \|u_m^0\| = \max_m |\psi(x_m)| \leqslant \|f^{(h)}\|_{F_h}.$$

Pour démontrer la stabilité qui a lieu pour $\tau \leqslant h$, il suffit de montrer que

$\|R_h^p\| \leqslant 1$. La validité de cette assertion découle de l'estimation $\|R_h\| \leqslant 1$:

$$\|R_h a\| = \max_m |a_m(1-r)+ra_{m+1}| \leqslant \max_m |a_m| = \|a\|.$$

EXEMPLE 2. Abordons maintenant un exemple plus compliqué et prenons à cet effet un autre schéma aux différences pour le même problème aux limites différentiel (10) :

$$\left.\begin{aligned}
&\frac{u_m^{p+1}-u_m^p}{\tau} - \frac{u_{m+1}^p-u_{m-1}^p}{2h} - \frac{\tau}{2}\,\frac{u_{m+1}^p-2u_m^p+u_{m-1}^p}{h^2} = \varphi_m^p,\\
&p = 0, 1, \ldots, [T/\tau]-1; \quad m = 1, 2, \ldots, M-1,\\
&u_m^0 = \psi(x_m), \quad m = 0, 1, \ldots, M,\\
&u_M^{p+1} = \psi_1(t_{p+1}), \quad p = 0, 1, \ldots, [T/\tau]-1,\\
&\frac{u_0^{p+1}-u_0^p}{\tau} - \frac{u_1^p-u_0^p}{h} = \varphi_0^p, \quad p = 0, 1, \ldots, [T/\tau]-1.
\end{aligned}\right\} \tag{12}$$

L'équation aux différences y figurant est d'ordre deux en x et l'équation différentielle correspondante (10) d'ordre un. Aussi impose-t-on à la frontière gauche $x = 0$ ($m = 0$) la condition supplémentaire

$$\frac{u_0^{p+1}-u_0^p}{\tau} - \frac{u_1^p-u_0^p}{h} = \varphi_0^p$$

que nous utiliserons sous la forme

$$u_0^{p+1} = (1-r)u_0^p+ru_1^p+\tau\varphi_0^p.$$

Nous avons examiné le schéma aux différences (12) au §23, où il s'agissait de l'approximation. Nous y avons muni l'espace F_h d'une norme de la façon suivante : si

$$f^{(h)} = \begin{cases}
a_m \text{ aux points } (mh, 0), & m = 0, 1, \ldots, M,\\
b^p \text{ aux points } (1, p\tau), & p = 0, 1, \ldots, [T/\tau],\\
c^p \text{ aux points } (0, p\tau), & p = 0, 1, \ldots, [T/\tau]-1,\\
\varphi_m^p \text{ aux points } (mh, p\tau), & m = 1, 2, \ldots, M-1;\\
 & p = 0, 1, \ldots, [T/\tau]-1,
\end{cases}$$

alors

$$\|f^{(h)}\|_{F_h} = h\max_p |c^p| + h\max_p \left|\frac{c^{p+1}-c^p}{\tau}\right| +$$
$$+\left(h\sum_{m=0}^{M}|a_m|^2\right)^{1/2} + \max_p |b^p| + \max_p \left|\frac{b^{p+1}-b^p}{\tau}\right| + \max_p \max_m |\varphi_m^p|.$$

Dans ce cas l'approximation est de l'ordre de h^2 (§23). Montrons qu'en choisissant la norme $\|u^{(h)}\|_{U_h}$ par la formule

$$\|u^{(h)}\|_{U_h} = \max_p \left(\frac{h}{2}|u_0^p|^2 + h\sum_{m=1}^{M}|u_m^p|^2\right)^{1/2},$$

on a pour $r \leqslant 1$ la stabilité en plus de l'approximation.

Vérifions la stabilité en ramenant le schéma aux différences (12) à la forme canonique (2). Posons

$$u^p = (u_0^p, u_1^p, \ldots, u_M^p)$$

de norme

$$\|u^p\| = \left(\frac{h}{2}|u_0^p|^2 + h\sum_{m=1}^{M}|u_m^p|^2\right)^{1/2}.$$

Définissons l'opérateur R_h par les formules suivantes :

Si $a = (a_0, a_1, \ldots, a_M)$, $b = (b_0, b_1, \ldots, b_M)$ et $b = R_h a$, alors

$$\left.\begin{aligned} b_0 &= (1-r)a_0 + ra_1, \\ b_m &= \left(-\frac{r}{2}+\frac{r^2}{2}\right)a_{m-1} + (1-r^2)a_m + \left(\frac{r}{2}+\frac{r^2}{2}\right)a_{m+1}, \\ &\qquad\qquad m = 1, 2, \ldots, M-1, \\ b_M &= 0. \end{aligned}\right\} \tag{13}$$

Dans ce cas

$$\varrho^p = \left[\varphi_0^p, \varphi_1^p, \ldots, \varphi_{M-1}^p, \frac{\varphi_1(t_{p+1})}{\tau}\right].$$

La condition 1° n'est évidemment pas remplie pour notre choix des normes. En effet, si par exemple

$$f^{(h)} = \begin{cases} 0 \text{ aux points } (mh, 0), & m = 0, 1, \ldots, M, \\ 0 \text{ aux points } (1, p\tau), & p = 0, 1, \ldots, [T/\tau], \\ 1 \text{ aux points } (0, p\tau), & p = 0, 1, \ldots, [T/\tau]-1, \\ 0 \text{ aux points } (mh, p\tau), & m = 1, 2, \ldots, M-1, \\ & p = 0, 1, \ldots, [T/\tau]-1, \end{cases}$$

alors

$$\varrho^p = (1, 0, \ldots, 0), \quad \|\varrho^p\| = \frac{\sqrt{h}}{\sqrt{2}}, \quad \|f^{(h)}\|_{F_h} = h,$$

si bien qu'aucun K_1 ne vérifie pour tous les h l'inégalité

$$\|\varrho^p\| \leqslant K_1 \|f^{(h)}\|_{F_h}.$$

Pour l'espace U_h' composé de vecteurs $u^p = (u_0^p, u_1^p, \ldots, u_M^p)$ et pour le choix mentionné des normes, il n'existe visiblement pas de R_h tel que la condition 1° soit satisfaite, mais il n'en est pas ainsi en ce qui concerne la condition 1*.

Avant de démontrer la dernière affirmation notons qu'avec une norme $\|f^{(h)}\|_{F_h}$ obtenue en posant

$$\|f^{(h)}\|_{F_h} = \sqrt{h}\max|c^p| + \left(h\sum_{m=0}^{M}|a_m|^2\right)^{1/2} + \max\left|\frac{b^p}{\tau}\right| + \max_p \max_m |\varphi_m^p|,$$

l'opérateur R_h défini par les formules (13) vérifie la condition 1°, mais l'approximation sera non pas d'ordre 2 en h, mais seulement de $^3/_2$. Nous savons réduire le problème aux limites (12) à la forme canonique (2) sans changer de normes et dans les conditions 1° et 2° si on prend pour U_h' l'ensemble des fonctions vectorielles discrètes

$$u^p = \begin{bmatrix} u_m^{p+1} \\ u_m^p \end{bmatrix}, \quad m = 0, 1, \ldots, M.$$

Ceci étant, on voit se compliquer la construction de R_h et l'évaluation des normes de ses puissances. Nous nous abstiendrons donc d'une telle réduction.

Montrons qu'avec le choix fait (13) de l'opérateur R_h il existe des z^p vérifiant la condition 1* :

$$\left\| \varrho^p - \frac{1}{\tau}\left(z^{p+1} - R_h z^p\right) \right\| < K_1 \|f^{(h)}\|_{F_h}.$$

Construisons la fonction $z^{(h)}$ comme dans l'exemple 1 et écrivons la solution stationnaire (indépendante de p) du problème

$$\left.\begin{aligned} &\frac{z_m^{p+1} - z_m^p}{\tau} - \frac{z_{m+1}^p - z_{m-1}^p}{2h} - \frac{\tau}{2}\,\frac{z_{m+1}^p - 2z_m^p + z_{m-1}^p}{h^2} = 0, \\ &\qquad m = 1, 2, \ldots, M-1, \\ &\frac{z_0^{p+1} - z_0^p}{\tau} - \frac{z_1^p - z_0^p}{h} = \varphi_0^p, \\ &\qquad z_M^p = \psi_1^p \end{aligned}\right\} \tag{14}$$

sous l'hypothèse de φ_0^p et ψ_1^p fixés et indépendants de p. Cette solution est de la forme

$$z_m^p = \varphi_0^p h \frac{1+r}{2}\left[\left(\frac{r-1}{r+1}\right)^m - \left(\frac{r-1}{r+1}\right)^M\right] + \psi_1^p.$$

La fonction $\{z_m^p\}$ vérifie l'estimation

$$\|z^p\| < \bar{K}\|f^{(h)}\|_{F_h}, \quad \bar{K} = 2.$$

En effet,

$$\|z^p\| = \left(\frac{h}{2}|z_0^p|^2 + h\sum_{m=1}^{M}|z_m^p|^2\right)^{1/2} <$$

$$< 2h \max_m |\varphi_m^p| + (1+h)\max_p |\psi_1(t_p)| < 2\|f^{(h)}\|_{F_h}.$$

Soit

$$\xi^p \equiv z^{p+1} - R_h z^p.$$

Puisque z^p est solution d'un problème stationnaire, on écrit

$$z^p = R_h z^p + \tau\delta^p,$$

où

$$\delta^p = \left(\varphi_0^p, 0, 0, \ldots, 0, \frac{\psi_1(t_p)}{\tau}\right).$$

Donc

$$\xi^p = z^{p+1} - z^p + \tau\delta^p,$$

de sorte que

$$\xi_m^p = \begin{cases} h(\varphi_0^{p+1} - \varphi_0^p)\dfrac{1+r}{2}\left[1 - \left(\dfrac{r-1}{r+1}\right)^M\right] + (\psi_1(t_{p+1}) - \psi_1(t_p)) + \tau\varphi_0^p, & m = 0, \\ h(\varphi_0^{p+1} - \varphi_0^p)\dfrac{1+r}{2}\left[\left(\dfrac{r-1}{r+1}\right)^m - \left(\dfrac{r-1}{r+1}\right)^M\right] + (\psi_1(t_{p+1}) - \psi_1(t_p)), & \\ & 1 < m < M, \\ 0, \quad m = M. & \end{cases}$$

Les coordonnées du vecteur $\tilde{\varrho}^p = \varrho^p - \dfrac{1}{\tau}\xi^p$ sont, par conséquent, de la forme

$$\tilde{\varrho}_m^p = \begin{cases} -\dfrac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau} - \dfrac{h}{\tau}(\varphi_0^{p+1} - \varphi_0^p)\dfrac{1+r}{2}\left[1 - \left(\dfrac{r-1}{r+1}\right)^M\right], & m = 0, \\ -\dfrac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau} - \dfrac{h}{\tau}(\varphi_0^{p+1} - \varphi_0^p)\dfrac{1+r}{2}\left[\left(\dfrac{r-1}{r+1}\right)^m - \left(\dfrac{r-1}{r+1}\right)^M\right] + \varphi_m^p, & \\ & 0 < m < M, \\ 0, \; m = M. & \end{cases}$$

L'inégalité (= la condition 1*)

$$\|\tilde{\varrho}^p\| = \|\varrho^p - \frac{1}{\tau}(z^{p+1} - R_h z^p)\| \leqslant K_1 \|f^{(h)}\|_{F_h}$$

est satisfaite :

$$\|\tilde{\varrho}^p\| = \left(\frac{h}{2}|\tilde{\varrho}_0^p|^2 + h\sum_{m=1}^{M-1}|\tilde{\varrho}_m^p|^2\right)^{1/2} \leqslant \left|\frac{\psi_1(t_{p+1}) - \psi_1(t_p)}{\tau}\right| + \frac{h|\varphi_0^{p+1} - \varphi_0^p|}{\tau} + \\ + \max_m |\varphi_m^p| \leqslant \|f^{(h)}\|_{F_h}.$$

En ce qui concerne la condition 2°

$$\|u^0\| \leqslant K_2 \|f^{(h)}\|_{F_h},$$

elle est évidente :

$$\|u^0\| = \left(\frac{h}{2}|u_0^2|^2 + h\sum_{m=1}^{M}|u_m^0|^2\right)^{1/2} \leqslant \left(h\sum_{m=0}^{M}|u_m^0|^2\right)^{1/2} \leqslant \|f^{(h)}\|_{F_h}.$$

On termine la démonstration de la stabilité ayant lieu pour $r \leqslant 1$ en montrant que dans cette condition on a

$$\|R_h^p\| \leqslant K, \quad p = 1, 2, \ldots, [T/\tau], \tag{15}$$

avec K une constante indépendante de h. Démontrons d'abord que tout vecteur $u = (u_0, u_1, \ldots, u_M)$ de composante u_M nulle vérifie l'inégalité

$$\|R_h u\| \leqslant \|u\|. \tag{16}$$

En faisant agir R_h sur le vecteur $(0, 0, \ldots, 0, 1)$ on obtient ensuite le vecteur $\left(0, 0, \ldots, 0, \frac{r}{2}+\frac{r^2}{2}, 0\right)$ de norme au plus égale à $\sqrt{h}$. On a donc pour un vecteur $u = (u_0, u_1, \ldots, u_M)$ arbitraire de composante u_M pas forcément nulle, compte tenu de (16) vrai pour un vecteur u de la forme $u = (u_0, u_1, \ldots, u_{M-1}, 0)$,

$$\begin{aligned} \|R_h u\| &= \|R_h(u_0, u_1, \ldots, u_{M-1}, 0)+u_M R_h(0, 0, \ldots, 0, 1)\| \leqslant \\ &\leqslant \|R_h(u_0, u_1, \ldots, u_{M-1}, 0)\|+|u_M|\sqrt{h} \leqslant \\ &\leqslant \|(u_0, u_1, \ldots, u_{M-1}, 0)\|+\|(0, 0, \ldots, u_M)\| \leqslant 2\|u\|, \\ &\|Rh\| \leqslant 2. \end{aligned} \tag{17}$$

Démontrons maintenant l'inégalité (15). Par définition de l'opérateur R_h, le vecteur $v = R_h u$ possède une composante nulle v_M. En utilisant (16) et (17) on trouve donc

$$\begin{aligned} \|R_h^p u\| &= \|R_h^{p-1}(R_h u)\| = \|R_h^{p-1} v\| \leqslant \|v\| = \\ &= \|R_h u\| \leqslant 2\|u\|, \quad \|R_h^p\| \leqslant 2 \equiv K. \end{aligned}$$

Il reste à légitimer l'inégalité (16) dont nous sommes partis, i.e. à démontrer l'affirmation suivante.

Soit $u = (u_0, u_1, \ldots, u_{M-1}, 0)$ un vecteur quelconque de composante u_M nulle et soit $v \equiv R_h u$. Alors $\|v\| \leqslant \|u\|$, i.e.

$$\left(\frac{h}{2} v_0^2+h\sum_{m=1}^{M} v_m^2\right)^{1/2} \leqslant \left(\frac{h}{2} u_0^2+h\sum_{m=1}^{M} u_m^2\right)^{1/2}. \tag{18}$$

Rappelons qu'en vertu de la définition (13) de R_h

$$\left.\begin{aligned} v_0 &= (1-r)u_0+ru_1, \\ v_m &= \left(-\frac{r}{2}+\frac{r^2}{2}\right)u_{m-1}+(1-r^2)u_m+\left(\frac{r}{2}+\frac{r^2}{2}\right)u_{m+1}, \\ &\qquad\qquad m = 1, 2, \ldots, M-1, \\ v_M &= 0. \end{aligned}\right\}$$

Signalons l'inégalité

$$\begin{aligned} v_m^2 \leqslant &\left[\left(-\frac{r}{2}+\frac{r^2}{2}\right)u_{m-1}+(1-r^2)u_m+\left(\frac{r}{2}+\frac{r^2}{2}\right)u_{m+1}\right]^2+ \\ &+\frac{r^2(1-r^2)}{4}(u_{m-1}-2u_m+u_{m+1})^2 \equiv \frac{r^2(1-r)}{2}u_{m-1}^2+ \\ &+(1-r^2)u_m^2+\frac{r^2(1+r)}{2}u_{m+1}^2-r(1-r^2)u_{m-1}u_m+r(1-r^2)u_m u_{m+1} \end{aligned}$$

vraie pour $r \leqslant 1$, et l'identité évidente

$$\frac{r^2(1-r)}{2}+1-r^2+\frac{r^2(1+r)}{2} = 1.$$

Pour $r \leqslant 1$, il est à présent facile d'établir la validité de la suite suivante sans supposer $u_M = 0$

$$\frac{1}{2} v_0^2 + \sum_{m=1}^{M} v_m^2 \leqslant \frac{1}{2} [u_0(1-r)+u_1 r]^2 + \sum_{m=1}^{M-1} \Big[\frac{r^2(1-r)}{2} u_{m-1}^2 +$$

$$+(1-r^2)u_m^2 + \frac{r^2(1+r)}{2} u_{m+1}^2 - r(1-r^2)u_{m-1}u_m +$$

$$+r(1-r^2)u_m u_{m+1} \Big] = \frac{1}{2} [u_0(1-r)+u_1 r]^2 + \sum_{m=1}^{M} u_m^2 +$$

$$+ \Big[\frac{r^2(1-r)}{2} u_0^2 - \frac{r^2(1+r)}{2} u_1^2 - r(1-r)u_0 u_1 \Big] - \frac{r^2(1-r)}{2} u_{M-1}^2 =$$

$$= \Big[\frac{1}{2} u_0^2 + \sum_{m=1}^{M} u_m^2 \Big] - \frac{r}{2} \{u_0^2 + [(1-r)u_0 + r u_1]^2\} -$$

$$- \frac{r^2(1-r)}{2} u_{M-1}^2 \leqslant \frac{1}{2} u_0^2 + \sum_{m=1}^{M} u_m^2.$$

L'inégalité de l'énergie ainsi obtenue

$$\frac{1}{2} v_0^2 + \sum_{m=1}^{M} v_m^2 \leqslant \frac{1}{2} u_0^2 + \sum_{m=1}^{M} u_m^2$$

est plus forte que l'inégalité (18) à démontrer.

Ainsi, la stabilité du schéma (12) pour $r \leqslant 1$ est un fait établi. Pour $r > 1$ il n'y a de stabilité pour aucun choix raisonnable des normes parce que la condition nécessaire de stabilité de Courant, Friedrichs et Lewy se trouve violée.

§41. Certains procédés d'évaluation des normes des puissances d'opérateurs

Aux §§39, 40 nous avons montré que les schémas aux différences évolutifs

$$L_h u^{(h)} = f^{(h)} \tag{1}$$

sont en général réductibles à la forme

$$\left.\begin{aligned} &u^{p+1} = R_h u^p + \tau \varrho^p, \\ &u^0 \text{ donné}, \end{aligned}\right\} \tag{2}$$

de façon que la stabilité soit équivalente à la propriété des normes des puissances de l'opérateur de passage

$$\|R_h^p\| < K, \quad p = 1, 2, \ldots, [T/\tau], \tag{3}$$

d'être bornées uniformément en h.

Les conditions (3) étant équivalentes à la stabilité, tout procédé d'étude de la stabilité est en même temps celui de vérification des inégalités (3).

Nous allons exposer, sous l'angle de l'évaluation des normes des puissances d'opérateurs, certaines approches de la stabilité introduites dès le chapitre 8 et nous en éclairons certains aspects nouveaux.

1. Conditions spectrales nécessaires pour que $\|R_h^p\|$ soient bornés. Soient λ_h une valeur propre de l'opérateur R_h et $u^{(h)}$ un vecteur propre correspondant, $R_h u^{(h)} = \lambda_h u^{(h)}$. Alors

$$\|R_h^p u^{(h)}\| = |\lambda_h|^p \|u^{(h)}\|,$$

et donc $\|R_h^p\| \geqslant |\lambda_h|^p$. Puisque λ_h est quelconque, on a

$$\|R_h^p\| \geqslant [\max|\lambda_h|]^p, \quad p = 1, 2, \ldots, [T/\tau], \tag{4}$$

où $\max|\lambda_h|$ est la plus grande valeur propre de R_h. Il est évident (voir §15) en vertu de (4) que la condition (3) est remplie s'il existe un cercle

$$|\lambda| \leqslant 1 + c\tau \tag{5}$$

du plan complexe (c étant une constante indépendante de h), qui comprend toutes les valeurs propres de l'opérateur R_h.

Le fait de prendre pour λ_h non seulement les valeurs propres de R_h mais aussi tous ses points du spectre ne complique guère les raisonnements ci-dessus sans changer le résultat. Si U_h' est un espace de dimension finie, le spectre de R_h est indépendant du choix de la norme et ne comprend que les valeurs propres. Ce cas fort important apparaît naturellement lors de l'approximation des problèmes aux limites différentiels dans un domaine borné par des problèmes aux différences sur un réseau D_h de nœuds en nombre fini. La condition (5) est alors nécessaire pour avoir (3) quelle que soit la norme choisie. Si le critère spectral nécessaire de stabilité n'est pas satisfait, le problème est irrémédiablement instable : on ne saurait y remédier par aucun choix raisonnable des normes. Nous avons traité en détail la situation analogue pour des équations aux différences ordinaires (§15).

Mettons en évidence la relation entre le critère spectral de stabilité du problème de Cauchy aux différences (critère de von Neumann considéré au §25) et le critère spectral (5) pour que les normes des puissances de $\|R_h\|$ soient uniformément bornées (3). A cette fin utilisons par exemple le schéma aux différences

$$\left.\begin{aligned} \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} &= \varphi(x_m, t_p), \\ u_m^0 &= \psi(x_m), \\ m = 0, \pm 1, \ldots;\ p = 0, 1, \ldots, [T/\tau]-1, & \end{aligned}\right\} \tag{6}$$

approchant le problème de Cauchy

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \varphi(x, t), \quad -\infty < x < \infty,\ 0 < t < T,$$
$$u(x, 0) = \psi(x).$$

Plus haut (§25) nous avons étudié sa stabilité moyennant le critère de von Neumann.

Récrivons le schéma en question sous forme canonique (2) en définissant R_h, $v = R_h u$ et ϱ^p par les formules

$$\left.\begin{aligned} v_m &= (1-r)u_m + ru_{m+1}, \quad r = \tau/h, \\ \varrho^p &= \varphi^p\{\equiv \varphi(x_m, t_p)\}. \end{aligned}\right\}$$

La norme dans U'_h se définit par $||u|| = \sup\limits_m |u_m|$. Quel que soit α réel, les fonctions $u = \{u_m\} = \{e^{i\alpha m}\}$ appartiennent alors à U'_h et sont fonctions propres de l'opérateur R_h:

$$R_h u = (1-r)e^{i\alpha m} + re^{i\alpha(m+1)} = [(1-r)+re^{i\alpha}]e^{i\alpha m} = \lambda(\alpha)u,$$

où

$$\lambda(\alpha) = 1-r+re^{i\alpha} \tag{7}$$

est valeur propre. Le nombre $\lambda(\alpha)$ étant indépendant de τ, la condition de stabilité (5) se ramène à la condition $|\lambda(\alpha)| \leqslant 1$ remplie pour tous les α réels lorsque $r \leqslant 1$.

Comme montré au §25, s'agissant du problème de Cauchy relatif à des schémas aux différences à deux niveaux en une fonction discrète cherchée, la condition (5) est non seulement nécessaire, mais aussi suffisante pour avoir la stabilité si la norme est définie par l'égalité

$$||u|| = \left(h \sum_{m=-\infty}^{\infty} u_m^2\right)^{1/2}.$$

(Dans ce cas $\{e^{i\alpha m}\}$ ne sont pas éléments de l'espace U'_h et ne constituent donc pas des fonctions propres, mais les points (7) appartiennent quand même au spectre de l'opérateur R_h.)

2. Critère spectral de puissances bornées d'un opérateur auto-adjoint. Soit U'_h un espace vectoriel de dimension M de fonctions définies aux points P_1, P_2, ..., P_m d'un réseau (ce dernier peut faire partie d'un segment, du plan et de l'espace, peu importe). Supposons que U'_h est muni d'un produit scalaire désigné par (u, v) pour deux fonctions u, $v \in U'_h$ quelconques et que R_h est un opérateur uniformément borné en h par une constante c_1:

$$||R_h|| < c_1, \tag{8}$$

et applique U'_h sur un sous-espace $\tilde{U}'_h \subset U'_h$ de dimension $N \leqslant M$, R_h étant auto-adjoint sur $\tilde{U}'_h$, i.e. $(R_h u, v) = (u, R_h v)$ pour tout couple de fonctions u, $v \in \tilde{U}'_h$. On sait du cours d'algèbre linéaire qu'il existe alors dans $\tilde{U}'_h$ une base orthonormale

$$\psi^{(1)}, \psi^{(2)}, \ldots, \psi^{(N)} \tag{9}$$

formée de vecteurs propres de l'opérateur R_h. Désignons par $\lambda_1, \lambda_2, \ldots, \lambda_N$ les valeurs propres (réelles) correspondantes:

$$R_h\psi^{(k)} = \lambda_k\psi^{(k)}, \quad k = 1, 2, \ldots, N. \tag{10}$$

THÉORÈME 1. *L'estimation* (3) *a lieu si et seulement si on a l'inégalité*

$$\max_k |\lambda_k| \leqslant 1+c_2\tau, \quad c_2 = \text{const}. \tag{11}$$

DÉMONSTRATION. La *nécessité* est démontrée au n° 1. Démontrons la *suffisance*. Soit $u \in U'_h$. Décomposons le vecteur $R_h u = v \in \tilde{U}'_h$ dans la base (9) :

$$v = \sum \alpha_k \psi^{(k)}.$$

En vertu de (10) on a alors

$$R_h^p u = R_h^{p-1} v = \sum_{k=1}^{M} (\lambda_k^{p-1}\alpha_k)\psi^{(k)},$$

$$\|R_h^p u\| = \left(\sum_{k=1}^{M} |\alpha_k \lambda_k^{p-1}|^2\right)^{1/2} \leqslant$$

$$\leqslant \max_k |\lambda_k^{p-1}| \left(\sum_{k=1}^{M} \alpha_k^2\right)^{1/2} = \max_k |\lambda_k^{p-1}| \cdot \|v\|. \tag{12}$$

Vu que, conformément à (8),

$$\|v\| = \|R_h u\| \leqslant c_1 \|u\|$$

et compte tenu de la condition (11) déduisons (3) de l'estimation (12) :

$$\|R_h^p\| \leqslant c_1 \max_k |\lambda_k^{p-1}| \leqslant c_1(1+c_2\tau)^{p-1} \leqslant$$

$$\leqslant c_1(1+c_2\tau)^{T/\tau} < c_1 e^{c_2 T} = K.$$

Plus loin nous établirons certains critères sous lesquels R_h est auto-adjoint et indiquons des procédés d'évaluation des valeurs propres.

3. Critères de R_h auto-adjoint. Introduisons la notation

$$[u, v] = \frac{1}{M} \sum u(P_k)\, v(P_k) \tag{13}$$

et supposons que le produit scalaire dans U'_h est défini par l'égalité

$$(u, v) = [u, v]. \tag{14}$$

Soit ensuite R_h, $b = R_h a$, un opérateur défini par les formules

$$b(P_k) = \sum_{P_s} \alpha_{ks} a(P_s),$$

P_s et P_k parcourant tous les points du réseau. R_h est auto-adjoint si et seulement si

$$\alpha_{ks} \equiv \alpha_{sk}. \tag{15}$$

La numérotation des points étant supposée sans importance, ce critère signifie qu'en calculant $b(P)$ en un point quelconque P du réseau, la valeur $a(Q)$ en un autre point Q doit figurer dans l'expression de $b(P)$ avec le même coefficient que la valeur $a(P)$ dans l'expression de $b(Q)$.

Si certains points du réseau $\{P_k\}$ constituent un sous-ensemble Γ_h (frontière du domaine) et si l'opérateur R_h est défini par les formules

$$\left.\begin{aligned} b(P_k) &= \sum_{P_s} \alpha_{ks} a(P_s), \quad & P_k \notin \Gamma_h, \\ b(P_k) &= 0, & P_k \in \Gamma_h, \end{aligned}\right\} \tag{16}$$

pour le sous-espace $\tilde{U}_h'$ ces dernières équivalent à

$$\left.\begin{aligned} b(P_k) &= \sum_{P_s \notin \Gamma_h} \alpha_{ks} a(P_s) \quad & \text{si} \quad P_k \notin \Gamma_h, \\ b(P_k) &= 0 & \text{si} \quad P_k \in \Gamma_h. \end{aligned}\right\}$$

La condition pour que R_h soit auto-adjoint sur le sous-espace $\tilde{U}_h'$ s'exprime (on le constate sans peine) par les égalités

$$\alpha_{ks} = \alpha_{sk}, \quad P_s \notin \Gamma_h, \quad P_k \notin \Gamma_h. \tag{17}$$

Ainsi, l'opérateur $b = R_h a$,

$$\left.\begin{aligned} b_k &= (1-2r)a_k + r(a_{k-1} + a_{k+1}), \quad k = 1, 2, \ldots, M-1, \\ b_0 &= b_M = 0 \end{aligned}\right\}$$

apparu en réduisant à la forme canonique (2) l'analogue aux différences de l'équation de la chaleur sur un segment, vérifie la condition (17) et ne vérifie pas (15).

4. Evaluation des valeurs propres de l'opérateur R_h. Dans certains cas on peut connaître exactement les valeurs propres d'un opérateur (au §27 nous l'avons fait pour Λ_{xx} agissant sur des fonctions définies sur un segment maillé et s'annulant en ses extrémités, ainsi que pour $\Lambda_{xx}+\Lambda_{yy}$ opérant sur des fonctions définies sur un réseau carré et s'annulant sur le contour).

En ce qui concerne les opérateurs aux différences auto-adjoints on recourt aux méthodes variationnelles. On sait que dans ce cas

$$\min_{u' \in U_h'} \frac{(R_h u', u')}{(u', u')} = \lambda'_{\min}, \quad \max_{u' \in U_h'} \frac{(R_h u', u')}{(u', u')} = \lambda'_{\max}. \tag{18}$$

Supposons par exemple que $\Lambda_{xx}+\Lambda_{yy}$ agit sur des fonctions discrètes de U_h' définies non pas sur un carré, mais sur un domaine plus compliqué composé de carrés et s'annulant sur sa frontière. Superposons ce domaine sur un réseau carré suffisamment vaste et considérons l'opérateur $\Lambda_{xx}+\Lambda_{yy}$ agissant sur des fonctions de U_h'' définies sur ce dernier domaine et s'annulant sur sa frontière.

Prolongeons chaque fonction $u' \in U_h'$ en une fonction $u'' \in U_h''$ en la faisant identiquement nulle en tous les points du réseau carré en dehors du domaine initial. On voit aisément que, vu sa nullité sur la frontière de celui-ci, une telle fonction satisfait à l'égalité

$$\frac{(R_h u', u')}{(u', u')} = \frac{(R_h u'', u'')}{(u'', u'')}, \quad R_h = \Lambda_{xx}+\Lambda_{yy}.$$

On obtient donc par passage des formules (18) à

$$\min_{u'' \in U_h''} \frac{(R_h u'', u'')}{(u'', u'')} = \lambda''_{\min},$$

$$\max_{u'' \in U_h''} \frac{(R_h u'', u'')}{(u'', u'')} = \lambda''_{\max},$$

les nombres $\lambda''_{\min}$ et $\lambda''_{\max}$ évalués par

$$\lambda''_{\min} \leqslant \lambda'_{\min}, \quad \lambda'_{\max} \leqslant \lambda''_{\max}. \tag{19}$$

Or, pour un domaine carré, les valeurs propres sont connues et il en est donc de même de $\lambda''_{\min}$ et $\lambda''_{\max}$, et nous avons obtenu les estimations (19) des frontières du spectre de l'opérateur $\Lambda_{xx}+\Lambda_{yy}$ agissant sur les fonctions de U'_h définies dans le domaine de départ.

De nombreux cas se traitent par des méthodes variationnelles d'évaluation des valeurs propres, analogues à celles pour les équations différentielles. Ainsi, la première valeur propre du problème

$$\tilde{R}_h u^{(h)} = \lambda u^{(h)}, \quad u^{(h)}|_{\Gamma_h} = 0,$$

où Γ_h est la frontière du domaine D_h et on a aux points intérieurs

$$\tilde{R}_h u_{mn} \equiv \frac{1}{h^2} \left\{ a\left(x_m + \frac{h}{2}, y_n\right)(u_{m+1,n} - u_{mn}) - \right.$$
$$\left. - a\left(x_m - \frac{h}{2}, y_n\right)(u_{mn} - u_{m-1,n}) \right\} +$$
$$+ \frac{1}{h^2} \left\{ b\left(x_m, y_n + \frac{h}{2}\right)(u_{m,n+1} - u_{mn}) - b\left(x_m, y_n - \right.\right.$$
$$\left.\left. - \frac{h}{2}\right)(u_{mn} - u_{m,n-1}) \right\},$$
$$a(x, y) > 0, \quad b(x, y) > 0,$$

ne peut que diminuer si on remplace les coefficients variables $a(x, y)$, $b(x, y)$ par les coefficients constants

$$a = \max_{x,y} a(x, y), \quad b = \max_{x,y} b(x, y).$$

Cela se démontre comme le fait analogue pour les équations différentielles (voir [9]).

Dans le cas des coefficients constants on peut passer du domaine initial au domaine carré et obtenir des estimations pareilles à (19). Le calcul exact des valeurs propres de l'opérateur $a\Lambda_{xx}+b\Lambda_{yy}$ dans un domaine carré s'avère facile.

5. Choix du produit scalaire. Supposons que R_h, $v = R_h u$, est défini par l'égalité

$$B_h v = \tilde{A}_h u \tag{20}$$

et que pour un certain choix fixe du produit scalaire $(u, v) \equiv [u, v]$ (pas

nécessairement d'après les formules (13), (14)) les opérateurs $\tilde{A}_h$ et B_h sont auto-adjoints :

$$[\tilde{A}_h u, v] \equiv [u, \tilde{A}_h v], \quad [B_h u, v] \equiv [u, B_h v].$$

Soit ensuite $B_h > 0$:

$$[B_h u, u] > 0 \quad \text{si} \quad u \neq 0.$$

L'opérateur $R_h = B_h^{-1}\tilde{A}_h$ est alors auto-adjoint au sens du produit scalaire

$$(u, v)_{B_h} \equiv [B_h u, v]. \tag{21}$$

En effet,

$$(B_h^{-1}\tilde{A}_h u, v)_{B_h} = [B_h(B_h^{-1}\tilde{A}_h)u, v] = [\tilde{A}_h u, v] =$$
$$= [u, \tilde{A}_h v] = [B_h^{-1}B_h u, \tilde{A}_h v] =$$
$$= [B_h u, B_h^{-1}\tilde{A}_h v] = (u, B_h^{-1}\tilde{A}_h v)_{B_h}.$$

L'identité démontrée en u et v

$$(B_h^{-1}\tilde{A}_h u, v)_{B_h} \equiv (u, B_h^{-1}\tilde{A}_h v)_{B_h}$$

signifie justement que R_h est auto-adjoint.

En choisissant le produit scalaire par la formule (21) on peut donc utiliser le critère spectral du n° 2 sous lequel les normes des puissances d'opérateurs auto-adjoints sont bornées. On affirme notamment que R_h défini par l'égalité (20) possède des valeurs propres réelles λ_k et un système orthonormal complet de vecteurs propres $\psi^{(k)}$:

$$\lambda_k B_h \psi^{(k)} = \tilde{A}_h \psi^{(k)} \tag{22}$$

et que l'appartenance de toutes les valeurs propres au segment $-1 \leqslant \lambda \leqslant 1$ est nécessaire et suffisante pour qu'on ait les inégalités

$$\|R_h^p\|_{B_h} \leqslant 1, \tag{23}$$

avec la norme de l'opérateur définie à l'aide du produit scalaire (21).

6. Critères de stabilité de Samarski. A. Samarski a proposé ([12], [13]) la théorie de la stabilité d'une vaste classe de schémas aux différences dans un espace de Hilbert qui indique les conditions nécessaires et suffisantes de stabilité en termes des inégalités linéaires entre les coefficients opératoriels de ces schémas et d'autres résultats.

Nous allons nous borner à deux résultats de la théorie en question.

Soit U_h' un espace euclidien muni d'un produit scalaire $(u, v) \equiv [u, v]$ et soit R_h, $v = R_h u$, $u, v \in U_h'$, défini par l'égalité

$$B_h \frac{v-u}{\tau} + A_h u = 0, \tag{24}$$

A_h et B_h étant des opérateurs auto-adjoints et $B_h > 0$. Définissons la norme énergie $\|u\|_{B_h}$ dans U_h' en posant

$$\|u\|_{B_h}^2 = [B_h u, u] \equiv (u, u)_{B_h}. \tag{25}$$

On a alors le

THÉORÈME 2. *Les conditions*

$$0 \leqslant A_h \leqslant \frac{2}{\tau} B_h \tag{26}$$

sont nécessaires et suffisantes pour qu'on ait

$$\|R_h^p\| \leqslant 1, \quad p \geqslant 0. \tag{27}$$

DÉMONSTRATION. Définissons un opérateur auto-adjoint $\tilde{A}_h$ par l'égalité $\tilde{A}_h \equiv B_h - \tau A_h$. L'égalité (24) est alors équivalente à (20) et les conditions (26) aux conditions $-B_h \leqslant \tilde{A}_h \leqslant B_h$, i. e. à

$$-[B_h u, u] \leqslant [\tilde{A}_h u, u] \leqslant [B_h u, u]. \tag{28}$$

Nous savons (n° 5) que R_h est auto-adjoint au sens du produit scalaire (21) et l'affirmation du théorème équivaut à affirmer que toutes les valeurs propres λ_k de R_h se situent sur le segment $-1 \leqslant \lambda \leqslant 1$ si et seulement si on a les conditions (28). Démontrons la dernière affirmation.

Supposons que les conditions (28) sont satisfaites. Multiplions scalairement l'égalité (22) par une fonction propre $\psi^{(k)}$ de l'opérateur R_h. Nous obtenons $\lambda_k[B_k\psi^{(k)}, \psi^{(k)}] = [\tilde{A}_h\psi^{(k)}, \psi^{(k)}]$, d'où

$$|\lambda_k| = \left| \frac{[\tilde{A}_h\psi^{(k)}, \psi^{(k)}]}{[B_h\psi^{(k)}, \psi^{(k)}]} \right| \leqslant 1.$$

Inversement, soit $\max|\lambda_k| \leqslant 1$. Montrons qu'on a les conditions (28). Soit $u = \sum c_k\psi^{(k)}$ la décomposition d'un élément $u \in U_h'$ dans la base $\{\psi^{(k)}\}$ orthonormale au sens du produit scalaire (21). Alors

$$\begin{aligned}
|[\tilde{A}_h u, u]| &= |[\tilde{A}_h \textstyle\sum c_k\psi^{(k)}, \sum c_k\psi^{(k)}]| = \\
&= |[\textstyle\sum c_k\tilde{A}_h\psi^{(k)}, \sum c_k\psi^{(k)}]| = \\
&= |[\textstyle\sum c_k\lambda_k B_h\psi^{(k)}, \sum c_k\psi^{(k)}]| = |[B_h \sum c_k\lambda_k\psi^{(k)}, \sum c_k\psi^{(k)}]| = \\
&= |(\textstyle\sum c_k\lambda_k\psi^{(k)}, \sum c_k\psi^{(k)})_{B_h}| = \sum c_k^2 |\lambda_k| \leqslant \\
&\leqslant \textstyle\sum c_k^2 = (u, u)_{B_h} = [B_h u, u].
\end{aligned}$$

Il en résulte $[B_h u, u] \geqslant [\tilde{A}_h u, u]$, ce qui équivaut aux conditions (28). Le théorème est démontré.

Notons que vérifier les conditions (28) c'est vérifier la non-négativité de toutes les valeurs propres des opérateurs $B_h - \tilde{A}_h$ et $B_h + \tilde{A}_h$ auto-adjoints au sens du produit scalaire $[u, v]$.

Donnons, sans le démontrer, un autre critère de stabilité dans le cas des schémas (24) tels que $B_h > 0$, $A_h = A_h^* > 0$. Introduisons dans l'espace U_h' la norme énergie $\|u\|_{A_h}$ en posant $\|u\|_{A_h}^2 \equiv [A_h u, u]$.

THÉORÈME 3. *La condition* $B_h \geqslant \frac{\tau}{2} A_h$ *est nécessaire et suffisante pour avoir l'inégalité* $\|R_h\|_{A_h} \leqslant 1$.

Le théorème 3 figure au n° 4 du §1, chap. VI, du livre [12] et se démontre sans approche spectrale. Dans notre cas celle-ci s'avère inopérante parce que B_h n'est pas auto-adjoint.

Exercice

Soit l'opérateur R_h, $b = R_h a$, défini par les formules

$$\left.\begin{aligned} b_m &= (1-r)a_m + ra_{m+1}, \qquad m = 0, 1, \ldots, M-1, \\ b_M &= 0. \end{aligned}\right\}$$

Démontrer qu'il est impossible de définir un produit scalaire dans l'espace U_h' de fonctions discrètes $\{a_m\}$, $m = 0, 1, \ldots, M$, de façon que R_h devienne auto-adjoint.

CHAPITRE 13

CRITÈRE SPECTRAL DE STABILITÉ DES PROBLÈMES AUX LIMITES ÉVOLUTIFS NON AUTO-ADJOINTS

Nous montrerons dans ce chapitre que le spectre d'un opérateur R_h non auto-adjoint ne permet pas de juger de la stabilité du problème aux limites aux différences dans un domaine borné, nous introduirons la notion de spectre d'une *famille* d'opérateurs $\{R_h\}$ et considérerons la position spectrale de la stabilité, position qui reste raisonnable dans le cas des problèmes non auto-adjoints dans des domaines bornés. Nous donnerons un critère spectral de stabilité qui est nécessaire et presque suffisant.

§42. Spectre d'une famille d'opérateurs $\{R_h\}$

1. Nécessité de perfectionner le critère spectral de stabilité. Dans le chapitre précédent nous avons montré qu'on peut en général ramener un problème aux limites aux différences évolutif à la forme

$$\left.\begin{aligned} u^{p+1} &= R_h u^p + \tau \varrho^p, \\ u^0 &\text{ donné}, \end{aligned}\right\} \tag{1}$$

de façon que la stabilité sur l'intervalle temporel $0 \leqslant t \leqslant T$ soit équivalente à la propriété des normes des puissances de R_h d'être bornées uniformément en h, i.e. à l'estimation

$$\|R_h^p\| < K, \quad p = 1, 2, \ldots, [T/\tau], \tag{2}$$

où τ est le pas en temps, $\tau = \tau(h)$.

Nous avons établi que pour avoir la condition (1), i.e. la stabilité, il faut que les valeurs propres de l'opérateur R_h soient intérieures au cercle

$$|\lambda| < 1 + c\tau \tag{3}$$

du plan complexe. On a montré (§41) que si R_h est auto-adjoint, la condition (3) est non seulement nécessaire, mais encore suffisante pour que les normes des puissances de R_h soient uniformément bornées (2). Le même fait a été établi au §25 pour le problème de Cauchy aux différences à coefficients constants dans le cas de schémas aux différences à deux niveaux en une

fonction inconnue que l'opérateur de passage soit auto-adjoint ou non. Dans le cas général des problèmes aux limites aux différences non auto-adjoints dans des domaines bornés, le critère nécessaire (3) est cependant fort loin de celui suffisant ni adéquat à la borne uniforme (2) des normes $\|R_h^p\|$ des puissances de R_h.

L'exemple suivant est là pour le prouver.

EXEMPLE. Pour le problème aux limites aux différences

$$\left.\begin{aligned} &\frac{u_m^{p+1}-u_m^p}{\tau}-\frac{u_{m+1}^p-u_m^p}{h}=\varphi(x_m, t_p),\\ &u_M^p=0, \qquad p=0, 1, \ldots, [T/\tau],\\ &u_m^0=\psi(x_m), \quad m=0, 1, \ldots, M;\ Mh=1, \end{aligned}\right\} \tag{4}$$

approchant le problème

$$\begin{aligned} &u_t-u_x=\varphi(x, t),\\ &u(1, t)=0, \quad 0<x<1, \quad 0<t<T,\\ &u(x, 0)=\psi(x) \end{aligned}$$

et ramené à la forme canonique usuelle (1), l'opérateur R_h, $v=R_hu$, se définit par les formules

$$\begin{aligned} &v_m=(1-r)u_m+ru_{m+1}, \quad m=0, 1, \ldots, M-1,\\ &v_M=0, \quad r=\tau/h. \end{aligned}$$

Sa matrice est de la forme

$$R_h=\left\{\begin{matrix} 1-r & r & 0 \ldots & 0 & 0\\ 0 & 1-r & r \ldots & 0 & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & 0 \ldots & 1-r & r\\ 0 & 0 & 0 \ldots & 0 & 0 \end{matrix}\right\} \tag{5}$$

Le spectre de celle-ci est composé de ses valeurs propres, i.e. des racines de l'équation

$$\det(R_h-\lambda E)=0 \quad \text{ou} \quad -\lambda(1-r-\lambda)^M=0.$$

Les racines $\lambda=0$ et $\lambda=1-r$ de cette équation constituent le spectre de l'opérateur R_h quel que soit h. Ce spectre est dans le cercle unité $|\lambda|\leqslant 1$ pour $0<r\leqslant 2$. Or, pour $1<r\leqslant 2$ le schéma (4) ne vérifie pas la condition de Courant, Friedrichs et Lewy si bien qu'il n'y a de stabilité $\|R_h^p\|<K$ pour aucune norme raisonnable.

En effet, on peut montrer que si $T\geqslant 1$ et $\|u\|=\max\limits_m|u_m|$, on a l'inégalité

$$\max_{p=1, 2, \ldots, [T/\tau]}\|R_h^p\|\geqslant|1-2r|^{1/h}=\varrho^{1/h}. \tag{6}$$

$r > 1$ entraîne $\varrho > 1$, de sorte que, lorsque $h \to 0$ et $\tau = rh \to 0$, la quantité $\max_p \|R_h^p\|$ croît exponentiellement et la condition $\|R_h^p\| < K$ se trouve grossièrement violée. On démontre l'inégalité (6) en notant que si $u_m^0 = (-1)^m$, $m = 0, 1, \ldots, M$, les valeurs u_m^p de la fonction

$$u^p = R_h^p u^0, \quad p = 1, 2, \ldots, M, \quad \text{pour} \quad m = 0, 1, \ldots, M-p,$$

sont définies par la formule

$$u_m^p = (-1)^m (1-2r)^p, \quad m = 0, 1, \ldots, M-p.$$

Donc

$$\|R_h^p u^0\| \geqslant |1-2r|^p \|u^0\|, \quad p \leqslant M,$$

de sorte que pour ces p, $p = 1, 2, \ldots, M$,

$$\|R_h^p\| \geqslant |1-2r|^p, \quad p = 1, 2, \ldots, M = 1/h,$$

et l'inégalité (6) est démontrée.

Ainsi, nous avons établi que le critère spectral nécessaire (2) utilisant les valeurs propres des opérateurs R_h est trop grossier pour R_h non auto-adjoints: dans notre exemple il ne détecte pas l'instabilité pour $1 < r \leqslant 2$.

2. Spectre d'une famille d'opérateurs (définition). Soit R_h un opérateur linéaire défini sur un espace vectoriel normé U_h'.

Notons $\{R_h\}$ l'ensemble des opérateurs R_h pour toutes les valeurs h que prend le paramètre h caractérisant la finesse du réseau. Par la nature même des schémas aux différences le pas h du réseau peut prendre des valeurs positives aussi petites qu'on le veut.

Un nombre complexe λ sera appelé *point du spectre de la famille d'opérateurs* $\{R_h\}$ si, quels que soient h_0 et ε positifs, il existe un h, $h < h_0$, tel que l'inégalité

$$\|R_h u - \lambda u\| < \varepsilon \|u\|$$

admette une solution u, $u \in U_h'$.

L'ensemble de tous ces λ s'appelle *spectre de la famille d'opérateurs* $\{R_h\}$.

3. Condition nécessaire de stabilité.

THÉORÈME 1. *Si au moins un point λ_0 du spectre de la famille d'opérateurs $\{R_h\}$ est extérieur au cercle unité du plan complexe, de sorte que $|\lambda_0| > 1$, il n'existe alors pas de constante K commune à tous les h telle qu'on ait l'inégalité*

$$\|R_h^p\| < K, \tag{7}$$

p parcourant les valeurs entières de 0 à $p = p_0(h)$, avec $p_0(h) \to \infty$ pour $h \to 0$.

DÉMONSTRATION. Admettons qu'il n'existe pas de nombres $h_0 > 0$ et $c > 0$ tels que pour tous les $h < h_0$ on ait l'estimation

$$\|R_h\| < c. \tag{8}$$

Sous cette hypothèse l'affirmation du théorème est évidente. Il reste donc à démontrer le cas où ces $h_0 > 0$ et $c > 0$ existent.

Posons $|\lambda_0| = 1+\delta$, où λ_0 est le point du spectre pour lequel $|\lambda_0| > 1$. Donnons-nous un K quelconque et choisissons p et ε de façon à avoir

$$(1+\delta)^p > 2K,$$

$$1-(1+c+c^2+ \dots +c^{p-1})\varepsilon > \frac{1}{2}.$$

Par définition d'un point du spectre de la famille $\{R_h\}$ on indique des h positifs aussi petits qu'on le veut tels qu'il existe un vecteur $u \in U_h$, solution de l'inégalité

$$\|R_h u-\lambda_0 u\| < \varepsilon \|u\|. \tag{9}$$

Posons

$$R_h u = \lambda_0 u+z. \tag{10}$$

Il est clair que $\|z\| < \varepsilon \|u\|$. On déduit de (10)

$$R_h^p u = \lambda_0^p u+(\lambda_0^{p-1}z+\lambda_0^{p-2}R_h z+ \dots +R_h^{p-1}z).$$

Puisque $|\lambda_0| > 1$, on a

$$\|\lambda_0^{p-1}z+\lambda_0^{p-2}R_h z+ \dots +R_h^{p-1}z\| \leqslant$$
$$\leqslant |\lambda_0|^p(1+\|R_h\|+\|R_h^2\|+ \dots +\|R_h^{p-1}\|)\varepsilon \|u\|,$$

et donc

$$\|R_h^p u\| \geqslant |\lambda_0|^p [1-\varepsilon(1+c+c^2+ \dots +c^{p-1})] \|u\| >$$
$$> (1+\delta)^p \frac{1}{2} \|u\| \geqslant 2K\frac{1}{2} \|u\| = K\|u\|.$$

Ceci étant, h est supposé suffisamment petit pour que p soit inférieur à $p_0(h)$.

K étant arbitraire, nous venons de démontrer que *la condition nécessaire de l'estimation* $\|R_h^p\| < K$ *est la position de tous les points du spectre de la famille d'opérateurs* $\{R_h\}$ *à l'intérieur ou sur la circonférence du cercle unité* $|\lambda| \leqslant 1$.

4. Notion de spectre de la famille d'opérateurs $\{R_h\}$ (discussion). Commençons par noter l'analogie entre la définition d'un point du spectre de la famille $\{R_h\}$ et celle d'un point du spectre d'un opérateur R, qu'on trouve dans tout cours d'analyse fonctionnelle (nous prendrons pour R l'opérateur R_h pour un certain h fixe).

Un point λ du plan complexe s'appelle *point du spectre de l'opérateur* R_h si, quel que soit ε positif, l'inégalité

$$\|R_h u-\lambda u\| < \varepsilon \|u\|$$

possède une solution u de l'espace U_h', domaine de définition de R_h.

En comparant les définitions d'un point du spectre de la *famille d'opérateurs* $\{R_h\}$ et d'un point du spectre de R_h, il peut venir à l'esprit que le premier spectre est constitué des points du plan complexe qui s'obtiennent par passage à la limite pour $h \to 0$ à partir des points du second lorsque

$h \to 0$ par rapport à toutes les sous-suites possibles. Mais cette hypothèse est en général fausse.

Considérons un opérateur R_h, $v = R_h u$, défini par les égalités

$$\left.\begin{aligned} v_m &= (1-r)u_m + ru_{m+1}, \qquad & m &= 0, 1, \ldots, M-1, \\ v_M &= 0, & Mh &= 1. \end{aligned}\right\} \tag{11}$$

L'opérateur (11) agit dans un espace vectoriel de dimension $M+1$ et est décrit par la matrice (5). On sait que le spectre de (5) se compose de ses valeurs propres, i.e. des racines λ de l'équation $\det(R_h - \lambda E) = 0$. Ces valeurs propres calculées dès le n° 1 sont $\lambda = 0$ et $\lambda = 1-r$. Ainsi, le spectre de l'opérateur R_h, h quelconque, comprend deux points $\lambda = 0$ et $\lambda = 1-r$ indépendants de h. Or, le spectre de la famille d'opérateurs $\{R_h\}$ (voir § 43) se compose non seulement de ces deux points, comme on pourrait penser, mais aussi de tous les points du cercle $|\lambda - 1 + r| < r$ de rayon r et de centre au point $\lambda = 1-r$ (fig. 27, p. 222). Pour $r \leqslant 1$ ce spectre est dans le cercle unité $|\lambda| \leqslant 1$ et pour $r > 1$ ce critère spectral nécessaire de stabilité ne joue plus : l'inégalité $\|R_h^p\| < K$ n'est pas vérifiée uniformément en h.

La figure 48 visualise la relation entre $\|R_h^p\|$ et $p\tau = prh$, $r = {}^3/_2$, pour différents h. Dans ce cas, le spectre de chaque opérateur R_h est formé de

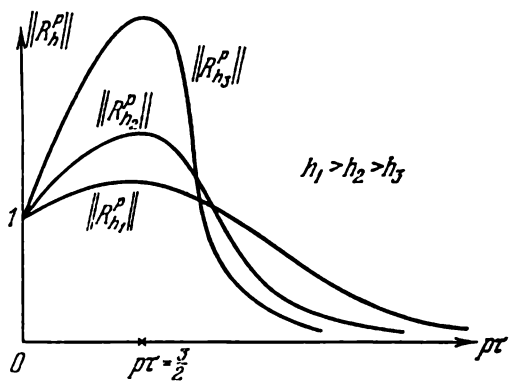


Fig. 48

deux points $\lambda = 0$ et $\lambda = -{}^1/_2$ intérieurs au cercle unité, ce qui explique l'allure de la courbe représentative de $\|R_h^p\|$ pour des valeurs importantes de $p\tau$. La quantité $\|R_h^p\|$ tend vers zéro pour $p\tau \to \infty$, i.e. l'axe des abscisses est une asymptote (on démontre dans les cours fondamentaux d'algèbre que la norme des puissances d'une matrice tend vers zéro avec la croissance de l'exposant si toutes les valeurs propres de la matrice sont inférieures à 1 en module).

Le fait que le spectre de la famille d'opérateurs $\{R_h\}$ n'est pas tout entier dans le cercle unité influe sur le comportement de $\|R_h^p\|$ pour $h \to 0$ et

des $p\tau$ pas très importants. La plus grande valeur de $\|R_h^p\|$ sur le segment $0 < p\tau < T$ (T est une constante positive arbitraire) croît vite avec la diminution de h. Mais cela signifie justement l'instabilité sur $0 < t < T$, et le comportement de $\|R_h^p\|$ pour $p\tau \to \infty$ lié au spectre de chaque opérateur R_h pris à part est inessentiel dans l'étude de la stabilité.

5. Le critère nécessaire de stabilité est proche d'un critère suffisant. Nous allons démontrer le

THÉORÈME 2. *Si un opérateur R_h est défini sur un espace normé U_h' de dimension finie pour chaque h et borné uniformément en h par une constante c :*

$$\|R_h\| < c \tag{12}$$

et que le spectre de la famille d'opérateurs $\{R_h\}$ se trouve tout entier dans le cercle unité fermé $|\lambda| \leqslant 1$, quel que soit $\varepsilon > 0$, les normes des puissances des opérateurs R_h sont alors évaluées par

$$\|R_h^p\| \leqslant A(\varepsilon)(1+\varepsilon)^p \tag{13}$$

avec $A = A(\varepsilon)$ dépendant de ε seul et non pas de h.

Cette affirmation signifie que non seulement la position du spectre de la famille $\{R_h\}$ dans le cercle unité est nécessaire pour qu'il y ait stabilité mais encore elle garantit contre une instabilité « grossière ». Sous les hypothèses du théorème la quantité

$$\max_{1 \leqslant p \leqslant [T/\tau]} \|R_h^p\|$$

reste bornée pour $h \to 0$ ou croît plus lentement que la puissance $\varrho^{[T/\tau]}$ de toute base $\varrho = 1+\varepsilon$ plus grande que un.

DÉMONSTRATION. Montrons au préalable que si le spectre de la famille d'opérateurs $\{R_h\}$ est dans le cercle $|\lambda| \leqslant \varrho$, il existe pour tout λ vérifiant $|\lambda| \geqslant \varrho+\varepsilon$, $\varepsilon > 0$, des nombres $A = A(\varepsilon)$ et $h_0 > 0$ tels que, quels que soient $h < h_0$ et $u \in U_h$, $u \neq 0$, on ait l'inégalité

$$\|R_h u - \lambda u\| > \frac{\varrho+\varepsilon}{A(\varepsilon)} \|u\|. \tag{14}$$

Supposons que le contraire est vrai. Il existe alors $\varepsilon > 0$, des suites de nombres $h_k > 0$, $h_k \to 0$; des suites de nombres complexes λ_k, $|\lambda_k| > \varrho+\varepsilon$, des suites de vecteurs $u_{h_k} \in U_{h_k}$ telles que

$$\|R_{h_k} u_{h_k} - \lambda_k u_{h_k}\| < \frac{\varrho+\varepsilon}{k} \|u_{h_k}\|. \tag{15}$$

Pour des k suffisamment grands tels que $\frac{\varrho+\varepsilon}{k} < 1$, les nombres λ_k ne peuvent être, en vertu de (12), extérieurs au cercle $|\lambda| \leqslant c+1$, car en dehors de celui-ci

$$\|R_{h_k} u - \lambda u\| \geqslant (|\lambda| - \|R_{h_k}\|) \|u\| \geqslant \|u\|.$$

La suite de λ_k est donc bornée et possède une valeur d'adhérence $\bar{\lambda}$, $|\bar{\lambda}| \geqslant \varrho+\varepsilon$. On constate sans peine qu'en vertu de (15) le point $\bar{\lambda}$ appartient au

spectre de la famille d'opérateurs $\{R_h\}$, ce qui contredit l'hypothèse sous laquelle ce spectre est dans le cercle $|\lambda| \leqslant \varrho$.

Soit maintenant R un opérateur linéaire d'un espace normé de dimension finie U dans lui-même. Supposons que tout λ complexe, $|\lambda| \geqslant r > 0$, et tout $u \in U$ pour $a = \text{const} > 0$ vérifient l'inégalité

$$\|Ru-\lambda u\| \geqslant a\|u\|. \tag{16}$$

Alors

$$\|R^p\| \leqslant \frac{r^{p+1}}{a}, \quad p = 1, 2, \ldots \tag{17}$$

L'inégalité (17) découle de l'égalité connue

$$R^p = -\frac{1}{2\pi i} \oint_{|\lambda|=r} \lambda^p (R-\lambda E)^{-1}\, d\lambda \tag{18}$$

et de la condition (16) qui veut que $\|(R-\lambda E)^{-1}\| \leqslant \frac{1}{a}$. On démontre l'inégalité (13) en posant $a = \frac{\varrho+\varepsilon}{A(\varepsilon)}$, $\varrho = 1$ et $R = R_h$. L'inégalité (17) coïncide alors avec (13).

Terminons ce paragraphe par une démonstration sommaire de l'égalité (18). Posons

$$u^{p+1} = Ru^p, \qquad U(\lambda) = \sum_{p=0}^{\infty} \frac{u^p}{\lambda^p},$$

multiplions les deux membres de $u^{p+1} = Ru^p$ par λ^{-p} et sommons sur p de $p = 0$ à $p = \infty$. Il vient

$$\lambda U(\lambda) - \lambda u^0 = RU(\lambda)$$

ou

$$(R-\lambda E)\, U(\lambda) = -\lambda u^0, \qquad U(\lambda) = -\lambda (R-\lambda E)^{-1} u^0.$$

La définition de $U(\lambda)$ montre que u^p est le résidu de la fonction vectorielle $\lambda^{p-1}U(\lambda)$:

$$u^p = \frac{1}{2\pi i} \oint_{|\lambda|=r} \lambda^{p-1} U(\lambda)\, d\lambda = -\frac{1}{2\pi i} \oint_{|\lambda|=r} \lambda^p (R-\lambda E)^{-1} u^0\, d\lambda.$$

Mais $u^p = R^p u^0$, de sorte que la dernière égalité est équivalente à l'égalité opératorielle (18).

Dans ce paragraphe nous avons formulé la position spectrale du problème de stabilité des schémas aux différences évolutifs, qui a un sens pour tout schéma de ce type réductible à la forme

$$\left.\begin{aligned} u^{p+1} &= R_h u^p + \tau \varrho^p, \\ u^0 &\text{ donné} \end{aligned}\right\}$$

de sorte que la condition

$$\|R_h^p\| < K, \quad p = 1, 2, \ldots, [T/\tau],$$

signifie la stabilité. Il peut s'agir en l'occurrence de schémas à deux ou à

plusieurs niveaux, de schémas de décomposition, etc., pour les problèmes sur un segment, dans des domaines multidimensionnels ou composés.

Cette position spectrale exige qu'on établisse si le spectre de la famille d'opérateurs $\{R_h\}$ se trouve dans le cercle unité $|\lambda| \leqslant 1$.

§ 43. Algorithme de calcul du spectre d'une famille d'opérateurs aux différences pour les fonctions définies sur un segment maillé

Nous allons décrire et justifier l'algorithme de calcul du spectre d'une famille d'opérateurs aux différences $\{R_h\}$ sur les espaces de fonctions discrètes (resp. de fonctions vectorielles) sur un segment. On accepte pour norme d'une fonction discrète (resp. d'une fonction vectorielle) le maximum des valeurs absolues de la fonction (resp. des composantes de la fonction vectorielle).

1. Exemple caractéristique. Soit $\{R_h\}$, $v = R_h u$, une famille d'opérateurs définie par les égalités

$$\left.\begin{aligned} v_m &= (1-r)u_m + ru_{m+1}, \quad m = 0, 1, \ldots, M-1, \\ v_M &= 0. \end{aligned}\right\} \tag{1}$$

L'opérateur R_h défini par (1) s'introduit dans la réduction naturelle du problème aux limites aux différences

$$\left.\begin{aligned} \frac{u_m^{p+1}-u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} &= \varphi(x_m, t_p), \quad p = 0, 1, \ldots, [T/\tau]-1, \\ u_M^{p+1} = 0, \quad u_m^0 &= \psi(x_m), \quad m = 0, 1, \ldots, M-1, \end{aligned}\right\} \tag{2}$$

à la forme

$$u^{p+1} = R_h u^p + \tau \varrho^p, \quad u^0 \text{ donné.}$$

Les relations (2) représentent l'analogue aux différences du problème aux limites différentiel

$$\begin{aligned} u_t - u_x &= \varphi(x, t), \quad 0 \leqslant x \leqslant 1, \quad 0 \leqslant t \leqslant T, \\ u(x, 0) &= \psi(x), \quad u(1, t) = 0. \end{aligned}$$

Nous avons déjà examiné le schéma aux différences (2) (n°2 du § 26) afin d'illustrer le critère de Babenko-Guelfand. On sait que selon ce critère l'étude d'un problème initial sur un segment se décompose en trois problèmes auxiliaires : un problème sans bords, un problème avec la seule frontière gauche, un problème avec la seule frontière droite, pour chacun desquels on doit trouver toutes les valeurs propres des opérateurs $u^p \to u^{p+1}$.

Il s'avère que l'algorithme de calcul du spectre de la famille $\{R_h\}$ coïncide avec le schéma de Babenko-Guelfand.

Pour pouvoir décrire l'algorithme en question on considère, en plus de R_h défini par les égalités (1), trois opérateurs auxiliaires $\bar{R}$, $\bar{R}$ et $\bar{R}$. L'opérateur $\bar{R}$, $v = \bar{R}u$, est défini sur un espace vectoriel de fonctions bornées

$u = \{\ldots, u_{-1}, u_0, u_1, \ldots\}$ définies sur toute la droite maillée $-\infty < mh < \infty$ par la formule

$$v_m = (1-r)u_m + ru_{m+1}, \quad m = 0, \pm 1, \ldots \tag{3}$$

Celle-ci s'obtient de (1) en éloignant la frontière gauche à $-\infty$ et la frontière droite à $+\infty$, ce qui se traduit par la flèche orientée dans les deux sens surlignant l'opérateur : $\bar{\bar{R}}$. L'opérateur $\bar{R}$, $v = \bar{R}u$, est défini sur un espace vectoriel de fonctions $u = (u_0, u_1, \ldots, u_m, \ldots)$ définies sur la demi-droite maillée $x_m = mh$, $m = 0, 1, 2, \ldots$, et tendant vers zéro pour $m \to +\infty$. Il est défini par la formule

$$v_m = (1-r)u_m + ru_{m+1}, \quad m = 0, 1, \ldots, \tag{4}$$

qui s'obtient de (1) par éloignement de la frontière droite à $+\infty$, fait exprimé par la flèche $\to$: $\bar{R}$.

Définissons enfin l'opérateur $\bar{R}$, $v = \bar{R}u$, agissant sur les fonctions

$$u = (\ldots, u_m, \ldots, u_{M-1}, u_M), \; u_m \to 0 \quad \text{pour} \quad m \to -\infty$$

définies sur la demi-droite maillée $x_m = mh$, $m = \ldots, -2, -1, 0, 1, \ldots$ $\ldots, M$, par les formules

$$\left.\begin{aligned} v_m &= (1-r)u_m + ru_{m+1}, \quad m = \ldots, -1, 0, 1, \ldots, M-1, \\ v_M &= 0. \end{aligned}\right\} \tag{5}$$

Les formules (5) s'obtiennent de (1) en éloignant la frontière gauche à $-\infty$, ce qui est également reflété dans la désignation de l'opérateur : $\bar{R}$.

Les opérateurs, $\bar{R}$, $\bar{\bar{R}}$ et $\bar{R}$ ne dépendent donc pas de h. Les domaines de définition des fonctions $u = \{u_m\}$ pour les opérateurs (1), (3), (4) et (5) sont représentés sur la figure 49. Nous démontrerons que *l'ensemble des valeurs propres des trois opérateurs constitue précisément le spectre de la famille d'opérateurs aux différences* $\{R_h\}$.

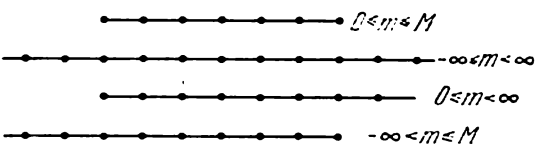


Fig. 49

Bien qu'ayant déjà calculé les valeurs propres de $\bar{\bar{R}}$, $\bar{R}$ et $\bar{R}$ (§ 26), nous répéterons l'opération parce qu'avant de passer à la démonstration même on doit se faire une idée nette de la structure des fonctions propres des opérateurs en question.

Commençons par établir l'ensemble des points λ du plan complexe pour lesquels l'équation

$$\bar{R}u - \lambda u = 0$$

possède une solution bornée $u = \{u_m\}$, $m = 0, \pm 1, \ldots$ Ces nombres λ

sont justement valeurs propres de l'opérateur $\bar{R}$. Dans notre exemple l'équation $\bar{R}u-\lambda u = 0$ est de la forme

$$(1-r-\lambda)u_m+ru_{m+1} = 0, \quad m = 0, \pm 1, \ldots$$

Toute solution de cette équation aux différences ordinaire du premier ordre ne diffère, comme cela découle du § 1, de la fonction discrète $u_m = q^m$, $m = 0, \pm 1, \ldots$, où q est racine de l'équation caractéristique $(1-r-\lambda)+ +rq = 0$, que par un facteur constant. La relation entre les nombres λ et q s'écrit également

$$\lambda = 1-r+rq.$$

La solution $u_m = q^m$ n'est bornée pour $m \to +\infty$ et $m \to -\infty$ que si $|q| = 1$, $q = e^{i\alpha}$, $0 \leqslant \alpha \leqslant 2\pi$. Aussi l'ensemble des λ tels que $u_m = q^m$ soit bornée s'obtient-il moyennant la formule

$$\lambda = 1-r+rq = 1-r+re^{i\alpha}$$

quand $q = e^{i\alpha}$ parcourt la circonférence unité $|q| = 1$ du plan complexe. Le point λ parcourt alors une circonférence $\bar{\Lambda}$ de rayon r et de centre au point $1-r$ (fig. 26, *a*, p. 222).

Calculons les valeurs propres de l'opérateur $\bar{R}$, i.e. des λ tels que l'équation

$$\bar{R}u-\lambda u = 0$$

admette une solution $u = (u_0, u_1, \ldots, u_m, \ldots)$ tendant vers zéro pour $m \to +\infty$.

Détaillons $\bar{R}u-\lambda u = 0$ comme suit :

$$(1-r-\lambda)u_m+ru_{m+1} = 0, \quad m = 0, 1, \ldots$$

Sa solution $u_m = q^m$, $m = 0, 1, \ldots$, tend vers zéro pour $m \to +\infty$ si $|q| < 1$. Les valeurs propres correspondantes $\lambda = 1-r+rq$ remplissent alors l'intérieur d'un cercle $\bar{\Lambda}$ de rayon r centré en $1-r$ (fig. 26, *b*).

L'algorithme de calcul des valeurs propres de l'opérateur $\bar{R}$ est analogue à l'algorithme précédent. Mettons $\bar{R}u-\lambda u = 0$ sous forme développée :

$$\left.\begin{aligned} (1-r-\lambda)u_m+ru_{m+1} &= 0, \quad m = \ldots, -1, 0, 1, \ldots, M-1, \\ -\lambda u_M &= 0. \end{aligned}\right\} \qquad (6)$$

Toute fonction discrète $u = \{u_m\}$, $m = M, M-1, \ldots$, vérifiant la première relation (6) est de la forme $u_m = q^m$ à un facteur constant près, et λ et q sont toujours reliés par l'égalité $\lambda = 1-r+rq$. La solution $u_m = = q^m$, $m = M, M-1, \ldots$, tend vers 0 pour $m \to -\infty$ si $|q| > 1$. La deuxième relation (6), i.e. $-\lambda u_M = 0$, impose à $u_m = q^m$ la condition supplémentaire $-\lambda u_M = -\lambda q^M = 0$ ou $\lambda = 0$. Si le point $\lambda = 0$ est extérieur au cercle de rayon r et de centre $1-r$ de la figure 26, *c*, i.e. si $r < 1/2$, il lui correspond une certaine valeur de q, $|q| > 1$. L'ensemble $\bar{\Lambda}$ des λ tels que l'équation $\bar{R}u-\lambda u = 0$ possède une solution tendant vers zéro pour $m \to -\infty$ se réduit au seul point $\lambda = 0$. Si $r \geqslant 1/2$, on constate, en vertu de l'analyse que nous venons de faire, que cette équation n'a pas de solution tendant vers 0 pour $m \to -\infty$ quel que soit λ complexe (ou réel).

La figure 27, *a* représente la réunion des valeurs propres des opérateurs $\bar{R}$, $\bar{\bar{R}}$ et $\bar{\bar{\bar{R}}}$ pour le cas $r < 1/2$ et les figures 27, *b* et 27, *c* pour le cas $r > 1/2$.

Démontrons que le spectre de la famille d'opérateurs $\{R_h\}$ coïncide avec la réunion Λ des ensembles $\bar{\Lambda}$, $\bar{\bar{\Lambda}}$, $\bar{\bar{\bar{\Lambda}}}$ des valeurs propres des opérateurs auxiliaires $\bar{R}$, $\bar{\bar{R}}$ et $\bar{\bar{\bar{R}}}$.

On doit montrer que tout point de Λ appartient au spectre de la famille d'opérateurs aux différences $\{R_h\}$ et que ce spectre ne possède pas d'autres points.

Montrons d'abord que tout point $\lambda_0 \in \Lambda$ appartient à ce spectre. Pour cela il suffit d'établir que, quel que soit $\varepsilon > 0$, l'inégalité

$$||R_h u - \lambda_0 u|| < \varepsilon ||u|| \tag{7}$$

a une solution u pour toutes les valeurs positives suffisamment petites de h. On peut dire de la solution $u = (u_0, u_1, \ldots, u_M)$ que c'est « quasi-vecteur propre » de l'opérateur R_h puisqu'en algèbre la solution de l'équation $R_h u - \lambda u = 0$ s'appelle vecteur propre.

Nos raisonnements dépendront de l'ensemble contenant le point λ_0. Commençons par le cas $\lambda_0 \in \bar{\Lambda}$ et montrons que pour tout $\varepsilon > 0$ et tous les h suffisamment petits, l'inégalité (7) admet u comme solution.

Construisons la fonction $u = (u_0, u_1, \ldots, u_M)$. Par définition de $\bar{\Lambda}$ il existe un q_0, $|q_0| = 1$ tel que $\lambda_0 = (1-r)+rq_0$ et l'équation $(1-r-\lambda_0)v_m + rv_{m+1} = 0$, $m = 0, \pm 1, \ldots$, possède une solution bornée $v_m = q_0^m$, $m = 0, \pm 1, \ldots$ Examinons cette solution pour $m = 0, 1, \ldots, M$ seuls tout en conservant la désignation v. Le vecteur

$$v = (v_0, v_1, \ldots, v_M) = (1, q_0, \ldots, q_0^M)$$

aurait évidemment satisfait à l'équation $R_h v - \lambda_0 v = 0$ qui s'écrit sous forme développée

$$\left.\begin{aligned} (1-r-\lambda_0)v_m + rv_{m+1} &= 0, \quad m = 0, 1, \ldots M-1, \\ -\lambda_0 v_M &= 0 \end{aligned}\right\}$$

si la dernière de ces relations n'avait pas été violée. La relation $-\lambda_0 v_M = 0$ peut s'interpréter comme la condition aux limites pour la solution de l'équation aux différences ordinaire

$$(1-r-\lambda_0)u_m + ru_{m+1} = 0, \quad m = 0, 1, \ldots, M-1.$$

On remplit cette condition aux limites donnée pour $m = M$, i.e. à l'extrémité droite du segment $0 \leqslant x \leqslant 1$, en « apportant une correction » au vecteur $v = (1, q_0, \ldots, q_0^M)$, c'est-à-dire en multipliant chacune des composantes v_m par $(M-m)h$. Le vecteur ainsi obtenu se note $u = (u_0, u_1, \ldots, u_M)$, $u_m = (M-m)hq_0^m$.

La figure 50 donne les courbes représentatives des fonctions $v = \{v_m\}$ et $u = \{u_m\}$ dans le cas $q_0 = -1$. La norme du vecteur u est égale à l'unité :

$$||u|| = \max_m |u_m| = \max_m |(M-m)hq_0^m| = Mh = 1.$$

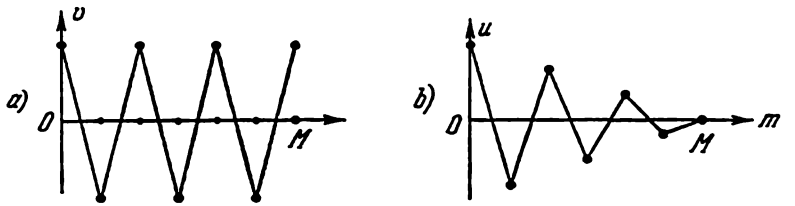


Fig. 50

Evaluons la norme du vecteur $w = (w_0, w_1, \ldots, w_M)$ défini par l'égalité $w \equiv R_h u - \lambda_0 u$. On a pour les coordonnées de w :

$$\begin{aligned} |w_m| &= |(1-r-\lambda_0)(M-m)hq_0^m + r(M-m-1)hq_0^{m+1}| = \\ &= |[(1-r-\lambda_0)+rq_0](M-m)hq_0^m - rhq_0^{m+1}| = \\ &= |0\cdot(M-m)hq_0^m - rhq_0^{m+1}| = rh, \quad m = 0, 1, \ldots, M-1, \\ |w_M| &= 0-\lambda_0\cdot 0 = 0. \end{aligned}$$

Cela entraîne $||w|| = rh$ et on a l'inégalité $||w|| = ||R_h u - \lambda_0 u|| < \varepsilon ||u||$ pour $h < \varepsilon/r$. Ainsi, nous avons démontré que le point $\lambda_0 \in \bar{\Lambda}$ appartient au spectre de la famille d'opérateurs aux différences $\{R_h\}$.

Montrons qu'il en est ainsi si λ_0 est élément de $\bar{\Lambda}$ ou de $\bar{\Lambda}$. Soit, pour fixer les idées, $\lambda_0 \in \bar{\Lambda}$. Par définition de l'ensemble $\bar{\Lambda}$ l'équation $\bar{R}v - \lambda_0 v = 0$ qui s'écrit sous forme développée

$$(1-r-\lambda_0)v_m + rv_{m+1} = 0, \quad m = 0, 1, 2, \ldots,$$

possède une solution $v_m = q_0^m$, $|q_0| < 1$, $m = 0, 1, \ldots$

Considérons cette solution seulement pour $m = 0, 1, \ldots, M$ en posant

$$u = (u_0, u_1, \ldots, u_M) = (1, q_0, \ldots, q_0^M).$$

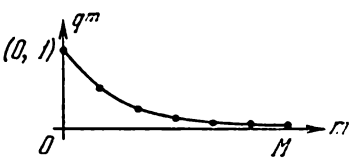


Fig. 51

Calculons pour cette fonction discrète u dont le graphe pour $q = 1/2$ est représenté sur la figure 51, la norme du vecteur $w \equiv R_h u - \lambda_0 u$. Les égalités

$$\begin{aligned} |w_m| &= |(1-r-\lambda_0)q_0^m + rq_0^{m+1}| = 0, \quad m = 0, 1, \ldots, M-1, \\ |w_M| &= |q_0|^M \end{aligned}$$

entraînent $||w|| = |q_0|^M = |q_0|^{1/h}$. Avec un h si petit que $|q_0|^{1/h} < \varepsilon$, on a $||w|| = ||R_h u - \lambda_0 u|| < \varepsilon = \varepsilon ||u||$ vu que $||u|| = 1$.

Ainsi, tous les points des ensembles $\bar{\Lambda}$, $\bar{\Lambda}$, $\bar{\Lambda}$ de notre exemple appartiennent au spectre de la famille d'opérateurs aux différences.

Montrons maintenant que tout point $\lambda_0 \in \bar{\Lambda}$, $\bar{\Lambda}$ et $\bar{\Lambda}$ n'appartient pas au spectre de la famille $\{R_h\}$, à savoir qu'il existe un nombre $A > 0$ indépendant de h et tel que toute fonction $u = (u_0, u_1, \ldots, u_M)$ vérifie l'inégalité

$$\|R_h u - \lambda_0 u\| \geqslant A \|u\|. \tag{8}$$

Quand $\varepsilon < A$, l'inégalité $\|R_h u - \lambda_0 u\| < \varepsilon \|u\|$ n'a alors aucune solution et le point λ_0 n'appartient pas au spectre. Notons $f \equiv R_h u - \lambda_0 u$. L'inégalité (8) prend alors la forme

$$\|f\| \geqslant A \|u\|. \tag{9}$$

C'est cette estimation que nous nous proposons de justifier. Mettons l'égalité $R_h u - \lambda_0 u = f$ sous forme développée

$$\left.\begin{aligned} (1-r-\lambda_0)u_m + ru_{m+1} &= f_m, \quad m = 0, 1, \ldots, M-1, \\ -\lambda_0 u_M &= f_M. \end{aligned}\right\} \tag{10}$$

Prenons ces relations pour équations en u et f pour un second membre donné. Ecrivons la solution $u = \{u_m\}$ comme une somme en posant

$$u_m = \alpha_m + \beta_m, \quad m = 0, 1, \ldots, M, \tag{11}$$

avec α_m les composantes d'une solution bornée $\alpha = \{\alpha_m\}$ de l'équation

$$(1-r-\lambda_0)\alpha_m + r\alpha_{m+1} = F_m = \begin{cases} 0 & \text{si} \quad m < 0, \\ f_m & \text{si} \quad m = 0, 1, \ldots, M-1, \\ 0 & \text{si} \quad m \geqslant M. \end{cases} \tag{12}$$

En vertu de la linéarité, le vecteur $\beta = \{\beta_m\}$ dont les composantes figurent dans (11) est solution de l'équation

$$\left.\begin{aligned} (1-r-\lambda_0)\beta_m + r\beta_{m+1} &= 0, \quad m = 0, 1, \ldots, M-1, \\ -\lambda_0 \beta_M &= f_M + \lambda_0 \alpha_M. \end{aligned}\right\} \tag{13}$$

Pour démontrer l'estimation (9) récrite, pour le choix ci-dessus de la norme, comme $|u_m| < \frac{1}{A} \max\limits_m |f_m|$, il suffit, en vertu de la relation $u_m = \alpha_m + \beta_m$, d'établir des estimations de la forme

$$|\alpha_m| \leqslant A_1 \max |f_m|, \tag{14}$$

$$|\beta_m| \leqslant A_2 \max |f_m|, \tag{15}$$

A_1 et A_2 étant des constantes. Commençons par (14) et notons que (12) est une équation du premier ordre de la forme

$$a\alpha_m + b\alpha_{m+1} = F_m, \quad m = 0, \pm 1, \ldots,$$

où $a = 1-r-\lambda_0$, $b = r$. Une équation de cette forme a été examinée au

§ 2 où nous avons abouti à l'estimation

$$|\alpha_m| \leqslant \frac{\max\limits_m |F_m|}{|a|-|b|} = \frac{\max\limits_m |f_m|}{|a|-|b|}. \tag{16}$$

Dans l'exemple considéré $|a|-|b| > \delta_0/2$, où δ_0 est la distance du point λ_0 à l'ensemble $\bar{\Lambda}+\bar{\Lambda}+\bar{\Lambda}$. L'estimation (16) entraîne donc l'inégaliét (14) que nous sommes en train de démontrer. L'estimation (15) découle de l'écriture suivante de la solution de (13) :

$$\beta_m = -\frac{f_M+\lambda_0\alpha_M}{\lambda_0}\, q_0^{m-M}, \tag{17}$$

avec q_0 défini par $(1-r-\lambda_0)+rq_0 = 0$. Par hypothèse, le point λ_0 n'est pas élément de Λ et il est donc extérieur à un cercle de centre en $1-r$ et de rayon r. Or, dans ce cas $|q_0| > 1$. Ensuite, $|\lambda_0| = \delta_1 > 0$ parce que pour $\lambda_0 = 0$ ce point appartiendrait à l'ensemble $\bar{\Lambda}+\bar{\Lambda}+\bar{\Lambda}$. Ainsi, en utilisant (17) et l'estimation (14) démontrée on obtient l'inégalité (15):

$$|\beta_m| = \left|\frac{f_M+\lambda_0\alpha_M}{\lambda_0}\right| \cdot |q_0^{m-M}| \leqslant \frac{|f_M|}{|\lambda_0|} + |\alpha_m| \leqslant$$
$$\leqslant \frac{\max|f_m|}{\delta_1} + A_1 \max|f_m| = A_2 \max|f_m|.$$

Nous venons donc de démontrer que le spectre de la famille d'opérateurs $\{R_h\}$ définie par la formule (1) coïncide avec la réunion des ensembles $\bar{\Lambda}$, $\bar{\Lambda}$ et $\bar{\Lambda}$ dans le plan complexe.

2. Algorithme de calcul du spectre dans le cas général.

THÉORÈME. *Etant donné un opérateur R_h, $b = R_h a$, $a, b \in U_h$, défini par l'égalité $B_h b = A_h a$, où A_h et B_h sont des opérateurs linéaires définis sur l'ensemble vectoriel normé de dimension finie U_h à valeurs dans un espace vectoriel normé F_h, uniformément bornés en h, B_h possédant un inverse $|B_h^{-1}|$ uniformément borné : $\|A_h\|$, $\|B_h\|$, $\|B_h^{-1}| < c$, le spectre de la famille d'opérateurs $\{R_h\}$ ne contient pas les λ du plan complexe pour lesquels l'opérateur $A_h - \lambda B_h$ admet, pour tous les h suffisamment petits, un inverse uniformément borné en h, et ces λ seulement.*

La démonstration étant évidente, nous nous en abstiendrons.

Soit maintenant R_h, $v = R_h u$, défini par les relations aux différences suivantes :

$$\left.\begin{aligned} &\sum_{k=-k_0}^{k_0} B_{km} v_{m+k} = \sum_{k=-k_0}^{k_0} A_{km} u_{k+m}, \quad k_0 \leqslant m \leqslant M-k_0, \\ &\sum_{i=0}^{2k_0} b_i v_i = \sum_{i=0}^{2k_0} a_i u_i, \qquad \sum_{i=0}^{2k_0} \beta_i v_{N-i} = \sum_{i=0}^{2k_0} \alpha_i u_{N-i}, \end{aligned}\right\} \tag{18}$$

le problème

$$\left.\begin{aligned} &\sum_{k=-k_0}^{k_0} B_{km} v_{m+k} = \varphi_m, \qquad k_0 \leqslant m \leqslant M-k_0, \\ &\sum_{i=0}^{2k_0} b_i v_i = \psi_1, \qquad \sum_{i=0}^{2k_0} \beta_i v_{M-i} = \psi_2 \end{aligned}\right\} \tag{19}$$

étant bien conditionné.

On suppose que

$$A_{km} = A_k\left(\frac{m}{M}\right), \quad B_{km} = B_k\left(\frac{m}{M}\right),$$

où $A_k(x)$ et $B_k(x)$ sont deux matrices carrées définies sur le segment $0 \leqslant x \leqslant 1$ et y vérifiant les conditions de régularité (14) § 4 ; a_i, b_i, α_i, β_i sont des matrices numériques rectangulaires indépendantes de M. Dans ce cas, on a le théorème ci-dessus et le spectre de la famille d'opérateurs $\{R_h\}$ est composé de tous les λ tels que le problème aux limites aux différences

$$\left.\begin{aligned} &\sum_{k=-k_0}^{k_0} (\lambda B_{km} - A_{km}) u_{m+k} = \varphi_m, \quad k_0 \leqslant m \leqslant M - k_0, \\ &\sum_{i=0}^{2k_0} (\lambda b_i - a_i) u_i = \psi_1, \quad \sum_{i=0}^{2k_0} (\lambda \beta_i - \alpha_i) u_{M-i} = \psi_2 \end{aligned}\right\} \tag{20}$$

ne soit pas bien conditionné. Pour voir si le problème (20) est bien conditionné, on se sert pour chaque λ du critère du n° 7 §4.

Exercices

1. Démontrer que pour la famille d'opérateurs aux différences $\{R_h\}$, $v = R_h u$, définie par les égalités

$$\left.\begin{aligned} &v_m = (1-r)u_m + r u_{m+1}, \quad m = 0, 1, \ldots, M-1, \\ &v_M = 0, \quad Mh = 1 \end{aligned}\right\}$$

et considérée dans le présent paragraphe, le spectre ne varie pas pour la norme définie non pas par la formule $\|u\| = \max_m |u_m|$, mais par $\|u\| = \left(h \sum_m |u_m|^2\right)^{1/2}$.

2. Démontrer que le spectre de la famille d'opérateurs aux différences $\{R_h\}$, $v = R_h u$, définie par les égalités

$$\left.\begin{aligned} &v_m = (1-r+\gamma h)u_m + r u_{m+1}, \quad m = 0, 1, \ldots, M-1, \\ &v_M = 0, \quad Mh = 1 \end{aligned}\right\}$$

ne dépend pas de la valeur de γ et coïncide avec le spectre construit dans ce paragraphe pour $\gamma = 0$.

3. Calculer le spectre de la famille d'opérateurs $\{R_h\}$, $v = R_h u$, définie par les égalités

$$\left.\begin{aligned} &v_m = (1-r)u_m + r(u_{m-1} + u_{m+1}), \quad m = 1, 2, \ldots, M-1, \\ &a u_0 + b u_1 = 0, \quad v_M = 0, \quad Mh = 1, \quad r = \text{const}, \end{aligned}\right\}$$

avec a et b donnés. Considérer les cas $|a| > |b|$ et $|a| < |b|$.

§ 44. Noyaux des spectres des familles d'opérateurs

Soit l'opérateur R_h d'un espace vectoriel normé U_h' d'une dimension N, $N = N(h)$, dans lui-même. Nous écrirons R_N et U_N au lieu de R_h et U_h' par souci de mettre en évidence la dimension. On suppose que $N \to \infty$ pour $h \to 0$.

Nous allons établir à quel point le spectre de la famille d'opérateurs $\{R_N\}$ dépend du choix de la suite de normes $\|\cdot\|_N$ dans les espaces U_N et, par là même, si le critère spectral de borne des normes des puissances de R_N (théorème 1 du § 42) est invariant par le choix des normes.

La famille $\{R_N\}$ sera supposée telle que les valeurs propres de tous les opérateurs R_N soient bornées dans l'ensemble, i.e. se trouvent dans un cercle

$$|\lambda| \leqslant c = \text{const}. \tag{1}$$

Il est évident que pour satisfaire à la condition (1) il suffit qu'il existe au moins une suite de normes $\|\cdot\|_N$ vérifiant les inégalités $\|R_N\|_N < c' = \text{const}$. D'où le caractère naturel de la contrainte (1) : elle est satisfaite pour les familles d'opérateurs R_N de passage d'un niveau à l'autre, qui apparaissent dans des problèmes aux limites aux différences évolutifs. Définissons à présent la notion de noyau du spectre qui nous permettra de formuler les résultats de ce paragraphe.

Soient donnés une suite de normes $\|\cdot\|_N$, un nombre $a \in [0, 1]$ et un $k \geqslant 0$ entier. Désignons par $\Lambda(a, k, N)$ l'ensemble des points λ tels que l'inégalité $\|R_N u - \lambda u\| \leqslant a^N N^{-k} \|u\|$ admette une solution $u \in U_N$, $u \neq 0$ (nous omettons l'indice N dans l'écriture des normes).

Définition. Appelons *noyau* de caractéristique $a \in [0, 1]$ de la famille d'opérateurs $\{R_N\}$ l'ensemble suivant $\Lambda(a)$ dans le plan complexe

$$\Lambda(a) \equiv \bigcap_{k \geqslant 0} \bigcap_{s \geqslant 0} \left(\overline{\bigcap_{N > s} \Lambda(a, k, N)} \right).$$

Ici

$$\overline{\bigcup_{N > s} \Lambda(a, k, N)} \equiv \Lambda_s(a, k)$$

est l'adhérence de la réunion des ensembles $\Lambda(a, k, N)$ pour tous les $N > s$;

$$\bigcap_{s \geqslant 0} \Lambda_s(a, k) = \Lambda(a, k)$$

l'intersection de tous les ensembles $\Lambda_s(a, k)$;

$$\bigcap_{k \geqslant 0} \Lambda(a, k) = \Lambda(a)$$

l'intersection de tous les ensembles $\Lambda(a, k)$.

Théorème 1. *Le noyau $\Lambda(a)$, $a \in [0, 1]$, est strictement contenu dans le spectre de la famille d'opérateurs $\{R_N\}$ et est fermé.*

Démonstration. Montrons que si le point λ_0 n'appartient pas au spectre de $\{R_N\}$, il n'est pas un point du noyau. En effet, il existe des $\varepsilon > 0$ et des N_0 tels que, pour tous les $N > N_0$, un $u \in U_N$ quelconque vérifie l'inégalité $\|R_N u - \lambda_0 u\| \geqslant \varepsilon \|u\|$. Tous les λ du cercle $|\lambda - \lambda_0| < \varepsilon/2$ vérifient alors l'inégalité $\|R_N u - \lambda u\| \geqslant \frac{\varepsilon}{2} \|u\|$. Ceci étant, pour $N > N_0$ aucun ensemble $\Lambda(a, k, N)$ n'a ses points dans le cercle $|\lambda - \lambda_0| < \varepsilon/2$. D'où la validité de la première affirmation du théorème. On démontre le caractère fermé du noyau $\Lambda(a)$ en remarquant que $\Lambda_s(a, k)$ le sont par construction et les ensembles $\Lambda(a, k)$ et $\Lambda(a)$ en tant qu'intersections d'ensembles fermés.

Exemple. Calculons le noyau $\Lambda(a)$, $a \in [0, 1]$, pour la famille d'opérateurs $\{R_N\}$ si l'opérateur R_{N+1}, $v = R_{N+1} u$, est défini par les égalités

$$\left.\begin{aligned} v_n &= (1-r)u_n + r u_{n+1}, \quad n = 0, 1, \ldots, N-1, \\ v_N &= 0, \end{aligned}\right\} \tag{2}$$

et la norme par $\|u\| = \|(u_0, u_1, \ldots, u_N)\| = \max |u_n|$.

Montrons que $\Lambda(a)$ se réduit au point $\lambda = 0$ et à un cercle fermé de rayon ar et de centre en $1 - r$:

$$|\lambda - (1-r)| \leqslant ar. \tag{3}$$

En effet, $\lambda = 0$ est, nous l'avons constaté au n° 1 du § 42, valeur propre de tous les opérateurs R_N et appartient par conséquent à tous les ensembles $\Lambda(a, k, N)$, donc au noyau. Tout λ_0 strictement intérieur au cercle (3) pour certains $\alpha \geqslant 0$ réel et $b > 1$ est

représenté par

$$\lambda_0 = 1 - r + \frac{ar}{b} e^{i\alpha}.$$

Pour tout k fixe et tous les N suffisamment grands, l'inégalité $\| R_N u - \lambda_0 u \| < a^N N^{-k} \| u \|$ a pour solution

$$u_n = \begin{cases} (a/b)^n e^{i\alpha n}, & n = 0, 1, \ldots, N-1, \\ 0, & n = N. \end{cases}$$

Par conséquent, les ensembles $\Lambda(a, k, N)$ contiennent, pour tous les N suffisamment importants, le point λ_0 qui est donc élément de $\Lambda(a)$. Ainsi, les points intérieurs au cercle (3) font partie du noyau $\Lambda(a)$ et, étant donné le caractère fermé de celui-ci, il comprend également la circonférence de (3).

Si le point $\lambda_0 \neq 0$ n'est pas dans le cercle (3), i. e.

$$\lambda_0 = 1 - r + \frac{ar}{b} e^{i\alpha}, \quad \alpha \geqslant 0, \ b = 1 - 2\delta, \ \delta > 0,$$

en écrivant la fonction de Green de l'équation aux différences du premier ordre (§ 2) on établit que pour n'importe quel λ du cercle $|\lambda - \lambda_0| < \min \left[|\lambda_0|, \frac{ar}{1-\delta} \right]$, tous les N suffisamment grands et tous les $u \in U_N$ on a l'inégalité $\| R_N u - \lambda u \| > a^N \| u \|$. Il en résulte que les points de ce cercle n'appartiennent pas à $\Lambda(a, k, N)$ si N est suffisamment grand et ne sont donc pas éléments de l'adhérence $\Lambda_s(a, k)$ de leur réunion ni du noyau $\Lambda(a)$.

Notons que le noyau $\Lambda(0)$, $a = 0$, de l'exemple considéré est formé de deux points $\lambda = 0$ et $\lambda = 1 - r$ et le noyau $\Lambda(1)$ coïncide avec le spectre tout entier de la famille $\{R_N\}$ que nous avons calculé au § 43.

Reprenons maintenant les constructions générales.

Définition. Nous appellerons le noyau $\Lambda(0)$ *noyau absolu.*

Théorème 2. *Le noyau absolu de la famille d'opérateurs $\{R_N\}$ ne dépend pas du choix de la suite de normes $\|\cdot\|_N$.*

La démonstration découle du fait que pour $a = 0$ l'ensemble $\Lambda(a, k, N)$ coïncide pour chaque N avec l'ensemble des valeurs propres de l'opérateur R_N, qui est indépendant de la norme dans l'espace U_N.

Théorème 3. *Dans la condition* (1) *on peut choisir une suite de normes $\|\cdot\|_N$ telle que le spectre de la famille d'opérateurs $\{R_N\}$ coïncide avec son noyau absolu.*

Démonstration. Construisons les normes dont le théorème affirme l'existence. Choisissons une base dans l'espace U_N de façon que la matrice de transformation R_N soit de Jordan dans cette base et que tous les termes extradiagonaux soient inférieurs à $1/N$ en module. Introduisons un produit scalaire et la norme associée en disant cette base orthonormale. Si λ_0 est un point quelconque ne faisant pas partie de $\Lambda(0)$ et $\varepsilon > 0$ la distance le séparant de l'ensemble $\Lambda(0)$ qui est fermé en vertu du théorème (1), on vérifie que $\| Ru - \lambda_0 u \| \geqslant \frac{\varepsilon}{4} \| u \|$ pour tous les $N > 8/\varepsilon$ et $u \in U_N$ si bien que λ_0 n'appartient pas au spectre de la famille d'opérateurs $\{R_N\}$.

Ainsi, si le spectre de la famille $\{R_N\}$ ne coïncide pas avec son noyau $\Lambda(0)$ de caractéristique $a = 0$ pour un choix donné des normes (cas de l'exemple (2) pour la norme $\|u\| = \max |u_n|$), on obtient alors pour spectre un ensemble $\Lambda(0)$ plus restreint à condition de choisir une autre suite de normes.

Les normes utilisées en théorie des schémas aux différences ne sont cependant pas tout à fait arbitraires.

Désignons par $\|\cdot\|_{cN}$ une norme égale au maximum des valeurs absolues de toutes les composantes d'une fonction discrète (ou d'une fonction vectorielle) $\in U_N$. Dégageons une classe de suites de normes $\|\cdot\|_N$ pour lesquelles il existe un s entier naturel fonction

de la suite et tel que pour tous les N suffisamment grands on ait l'inégalité

$$\sup_{\|u\|_e=1} \|u\|_N \leqslant N^s \inf_{\|u\|_e=1} \|u\|_N. \tag{4}$$

Il est évident que la norme $\|\cdot\|_{e_N}$ même et toutes les normes intervenues rapport aux équations aux différences forment une suite de classe (4) pour N croissant.

THÉORÈME 4. *Le noyau $\Lambda(a)$ de caractéristique $a \in [0, 1]$ ne dépend pas du choix de la suite de normes parmi celles vérifiant la condition* (4).

La DÉMONSTRATION résulte immédiatement des définitions.

Considérons maintenant la famille d'opérateurs R_k définis par les égalités (18) et (19) du § 43 sous l'hypothèse complémentaire que les coefficients matriciels A_k et B_k sont constants : $A_k(x) \equiv A_k(0)$, $B_k(x) \equiv B_k(0)$. On a pour cette famille le théorème important suivant :

THÉORÈME 5 (DE A. SOKOLOV). *Si on munit les espaces $U'_k = U_N$, où opèrent $R_k = R_N$, de normes $\|\cdot\|_{e_N}$, alors le noyau $\Lambda(1)$ de caractéristique $a = 1$ du spectre de la famille d'opérateurs $\{R_N\}$ coïncide avec tout le spectre de cette famille.*

Les deux derniers théorèmes entraînent que, quelle que soit la suite de normes appartenant à la classe vérifiant la condition (4), le spectre de la famille d'opérateurs $\{R_N\}$ contient le spectre de cette famille obtenu pour les normes $\|\cdot\|_{e_N}$, dont le calcul est décrit au n° 2 du § 43. Si la condition spectrale de borne des normes des puissances des R_N (théorème 1 du § 42) n'est pas remplie pour les normes $\|\cdot\|_{e_N}$, elle ne l'est non plus pour aucun autre choix de la suite des normes vérifiant (4).

La démonstration du théorème de Sokolov suppose une étude approfondie que nous laissons de côté.

§ 45. Sur la stabilité des algorithmes itératifs de résolution d'équations aux différences non auto-adjointes

La résolution par stationnarisation de problèmes stationnaires peut s'interpréter comme un processus itératif et les résultats obtenus à un niveau temporel comme l'approximation correspondante.

Au § 35 nous avons pris comme exemple le problème de Dirichlet aux différences pour l'équation de Poisson. Si les conditions aux limites sont nulles, il s'agit en l'occurrence d'un problème auto-adjoint, et on a donc pu, dans le processus de stationnarisation, décomposer l'erreur dans un système orthogonal complet de fonctions propres. La position des valeurs propres nous a permis de se faire une idée et de la vitesse de décroissance de l'erreur et de l'influence des erreurs d'arrondi faites aux niveaux intermédiaires.

Il se trouve que ce n'est en général pas le cas des équations aux différences non auto-adjointes qu'on résout par stationnarisation. Malgré la convergence du processus itératif, il peut y avoir l'instabilité due à une grande sensibilité aux erreurs d'arrondi.

Dans le présent paragraphe nous nous proposons de définir rigoureusement et de discuter le phénomène en question. Nous aurons besoin pour cela des notions de spectre et de noyau du spectre d'une famille d'opérateurs aux différences.

Soit

$$u = R_N u + f_N \tag{1}$$

une famille d'équations linéaires (« équation aux différences ») en un élément inconnu u d'un espace vectoriel normé U_N de dimension N, qui dépend d'un paramètre N, N entier naturel. Nous considérerons le processus itératif

$$u^{p+1} = R_N u^p + f_N, \quad p = 0, 1, \ldots, \tag{2}$$

de calcul de la solution u et supposerons que toutes les valeurs propres λ_k $\lambda_k(N)$ de

l'opérateur R_N ont un module inférieur à l'unité

$$|\lambda_k| < \varrho_N < 1, \tag{3}$$

i.e. on a le critère connu de convergence de (2) et

$$\|u-u^p\| = O(\varrho_N^p). \tag{4}$$

Effectuons les calculs (2) d'une façon approchée avec un nombre $p = q+\alpha$ de décimales, i.e. par la formule

$$\tilde{u}^{p+1} = R_N\tilde{u}^p+f_N+10^{-q}\|\tilde{u}^p\|\,\delta_p, \tag{5}$$

avec $\delta_p \in U_N$, $\|\delta_p\| \leqslant 1$ arbitraires.

Donnons-nous un s entier naturel et exigeons que pour n'importe quels δ_k, $\|\delta_k\| \leqslant 1$, $k = 0, 1, \ldots$, on ait l'inégalité

$$\overline{\lim_{p\to\infty}}\,\|u-\tilde{u}^p\| \leqslant 10^{-s}\|u\|. \tag{6}$$

Celle-ci garantit la possibilité de calculer la solution u par les formules (5) avec une erreur au plus égale à l'unité de la s-ième décimale (au sens de la norme dans U_N).

LEMME. *Pour avoir* (6) *il faut que le nombre* α *de décimales « de réserve » dans la formule* (5) *vérifie l'inégalité*

$$(1-10^{-s})\varphi \leqslant 10^{\alpha}$$

et il suffit que α *satisfasse à*

$$(1+10^{-s})\varphi \leqslant 10^{\alpha},$$

où

$$\varphi = \lim_{p\to\infty}\max_{\|\delta_k\|=1}\left\|\sum_{k=0}^{p} R_N^{p-k}\delta_k\right\|.$$

La démonstration du lemme est laissée au lecteur.

Notons que l'existence de $\varphi = \varphi(N)$ découle de la condition (3). Dans la suite nous entendrons par $\alpha = \alpha(N)$ le plus petit entier garantissant la condition (6). Il résulte du lemme qu'un tel nombre existe, est non négatif et dépend très peu de s ou n'en dépend pas du tout.

DÉFINITION. L'algorithme itératif convergent (2) sera dit *stable* s'il existe une constante C indépendante de N telle qu'on ait

$$\alpha(N) < C; \tag{7}$$

l'algorithme (2) sera dit *stable au sens faible* s'il existe une constante C indépendante de N telle qu'on ait

$$\alpha(N) < C\ln N, \tag{8}$$

sans qu'il y ait stabilité ; l'algorithme (2) sera enfin dit *instable* s'il n'est ni stable ni stable au sens faible.

EXEMPLE. Ecrivons l'équation

$$\left.\begin{aligned} -2u_n+u_{n+1}+f_n &= 0, \quad n = 0, 1, \ldots, N-1,\\ u_N &= 0 \end{aligned}\right\} \tag{9}$$

comme

$$\left.\begin{aligned} u_n &= (1-2r)u_n+ru_{n+1}+rf_n, \quad n = 0, 1, \ldots, N-1,\\ u_N &= 0, \end{aligned}\right\} \tag{10}$$

où $r > 0$ est un paramètre. L'algorithme itératif (2) s'écrit pour (10)

$$\left.\begin{aligned} u_n^{p+1} &= (1-2r)u_n^p+ru_{n+1}^p+rf_n, \quad n = 0, 1, \ldots, N-1,\\ u_N^{p+1} &= 0 \end{aligned}\right\} \tag{11}$$

de sorte que l'opérateur R_N, $v = R_N u$, se définit par les formules

$$v_n = (1-2r)u_n + ru_{n+1},$$
$$v_N = 0.$$

On constate aisément que R_N ne possède que deux valeurs propres $\lambda_1(N) = 1-2r$ et $\lambda_2(N) = 0$.

L'inégalité (3) a lieu et l'algorithme itératif (11) converge pour $r < 1$. Prenons pour norme $\|u\| = \max |u_n|$. Montrons que, lorsque $r < {}^2/_3$, l'algorithme (11) est stable et qu'il est instable pour $r > {}^2/_3$. En effet, si $r < {}^2/_3$, alors

$$\max_n |v_n| \leqslant \max(|1-3r|, |1-r|) \max_n |u_n|$$

si bien que $\|R_N\| \leqslant \max(|1-3r|, |1-r|) = \varrho < 1$. Aussi $\varphi(N) \leqslant 1/1-\varrho$ et on a, en vertu du lemme, l'estimation (7) où $C = -2\ln(1-\varrho)$. Soit maintenant $r > {}^2/_3$. Posons dans (11) $f_n \equiv 0$, $u_n^0 = (-1)^n$, $n = 0, 1, \ldots, N-1$. On voit sans peine que dans ce cas $u_n^p = (1-3r)^p(-1)^n$, $n = 0, 1, \ldots, N-p$. D'où $\|R_N^p\| \geqslant \varrho^p$, $p = 1, 2, \ldots, N-1$, où $\varrho = |1-3r| > 1$. Donc $\varphi(N) \geqslant \varrho^N$, et $\alpha \approx N \ln \varrho$ en vertu du lemme, ce qui démontre l'instabilité. On montre que quand $r = {}^2/_3$, il y a stabilité au sens faible.

Ainsi, le critère spectral de convergence (3) de l'algorithme itératif (2) ne détermine pas sa stabilité. Le critère spectral et les conditions de stabilité s'énoncent non pas en termes de la position des spectres de chacun des R_N, mais en ceux de la position du spectre et des noyaux du spectre de la famille $\{R_N\}$. Sous l'hypothèse de la famille de R_N uniformément bornée, $\|R_N\| < C$, on vérifie notamment sans difficulté les affirmations suivantes.

LEMME. *Pour que l'algorithme itératif* (2) *soit convergent pour tous les N suffisamment grands, il suffit que le rayon ϱ d'un noyau du spectre de la famille $\{R_N\}$ soit strictement inférieur à l'unité.*

CRITÈRE DE STABILITÉ. *L'algorithme itératif* (2) *est stable si et seulement si le spectre de la famille d'opérateurs $\{R_N\}$ soit strictement intérieur au cercle unité.*

THÉORÈME. *Pour que l'algorithme itératif* (2) *converge et soit stable ou stable au sens faible, il suffit que le rayon ϱ du noyau $\Lambda(1)$ du spectre de la famille d'opérateurs $\{R_N\}$ soit < 1 ; pour que l'algorithme itératif convergent* (2) *soit instable, il suffit que le rayon ϱ de ce noyau soit > 1.*

Nous avons montré au paragraphe précédent que le noyau $\Lambda(1)$ du spectre de la famille $\{R_N\}$ est indépendant du choix des normes de classe (4) naturelle aux équations aux différences. Il en découle en particulier que si les opérateurs R_N sont contractants uniformément en N, $\|R_N\| < C < 1$, de sorte que le spectre ainsi que le noyau $\Lambda(1)$ du spectre de $\{R_N\}$ se trouvent dans le cercle $|\lambda| < C$, alors l'algorithme itératif (2) est stable et la stabilité (forte ou faible) se conserve dans toute autre norme (4), § 44 pour laquelle les opérateurs R_N peuvent ne pas être contractants.

Dans l'exemple ci-dessus le spectre de la famille $\{R_N\}$ se compose du cercle $|\lambda-(1-2r)| < r$ et du point $\lambda = 0$ et il coïncide avec son noyau $\Lambda(1)$.

L'affirmation sur la stabilité de (11) pour $r < {}^2/_3$ et sur son instabilité lorsque $r > {}^2/_3$ se démontre donc également à l'aide des critères spectraux.

Pour calculer la solution d'une équation (non auto-adjointe) de la forme

$$A_N u + f_N = 0, \tag{12}$$

on essaie de construire un algorithme itératif de la forme

$$B_N u^{p+1} = B_N u^p + (A_N u^p + f_N). \tag{13}$$

On choisit alors B_N tel que son inversion numérique soit facile et que le spectre de la famille d'opérateurs $\{R_N\}$, $R_N \equiv E_N + B_N^{-1} A_N$, possède le plus petit rayon ϱ, $\varrho < 1$, possible. En vertu de l'estimation $\|R_N^p\| \leqslant C(\varepsilon)(\varrho+\varepsilon)^p$, où $\varepsilon > 0$ est arbitraire et $C(\varepsilon)$ indépendant de N (voir § 44), cela garantit une convergence rapide et, vu le critère de stabilité ci-dessus, la stabilité de l'algorithme itératif (13).

ANNEXE

MÉTHODE DES CONDITIONS AUX LIMITES INTRINSÈQUES

La théorie des problèmes aux limites relatifs aux fonctions analytiques, i.e. pour la résolution d'un système d'équations de Cauchy-Riemann, ainsi que pour la résolution de systèmes d'équations aux dérivées partielles plus généraux, fait appel à la méthode des équations intégrales singulières. Ce procédé consiste à ramener les problèmes aux limites à des équations intégrales sur la frontière du domaine considéré. En plus des conditions aux limites données on utilise alors les conséquences du système différentiel même, à savoir les relations que doivent vérifier les fonctions (et leurs dérivées normales) sur ladite frontière pour qu'on puisse les prolonger à l'intérieur du domaine en une solution du système correspondant. S'agissant des fonctions analytiques, c'est la condition classique de Sokhotski-Plemelj qui intervient en passant à la limite pour z tendant vers γ dans la formule intégrale de Cauchy

$$\varphi(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{\zeta - z} d\zeta.$$

Dans le cas d'équations différentielles du second ordre, la condition correspondante résulte de la formule de Green donnant la solution en chaque point du domaine en fonction de ses valeurs et des valeurs de sa dérivée normale sur la frontière. On obtient cette condition en passant à la limite lorsque le point tend vers la frontière de l'intérieur et en utilisant les propriétés des potentiels de simple et de double couche.

L'idée de la méthode des conditions aux limites intrinsèques est analogue à celle de la méthode de réduction des problèmes aux limites pour les équations aux dérivées partielles aux équations intégrales sur la frontière. Le rôle de conditions aux limites supplémentaires analogues à la condition de Sokhotski-Plemelj incombe à des conditions aux limites intrinsèques résultant de l'analogue aux différences de la formule intégrale de Cauchy (ou de celui de la formule de Green).

1. Classe de systèmes d'équations aux différences. On considère les problèmes aux limites pour des systèmes généraux d'équations aux différences à coefficients constants, qui s'écrivent sous forme vectorielle

$$\mathbf{L}u \equiv \sum_{k \in K} A_k u_{n+k} = f_n, \tag{1}$$

$n = (n_1, n_2, \ldots, n_s)$, $k = (k_1, k_2, \ldots, k_s)$ étant des multi-indices, A_k des matrices carrées, f_n une fonction vectorielle donnée et u_n la fonction vectorielle cherchée, K un ensemble fini (pochoir). Le système (1) est supposé vérifier la condition algébrique suivante : la matrice caractéristique

$$A(\xi) \equiv \sum_{k \in K} A_k \xi^{-k}, \tag{2}$$

où $\xi^{-k} \equiv \xi_1^{-k_1} \ldots \xi_s^{-k_s}$ et $\xi_1, \ldots, \xi_s$ sont des paramètres complexes, n'est pas dégénérée

identiquement en ξ :

$$\det A(\xi) \not\equiv 0. \tag{3}$$

Cette contrainte est naturelle : on montre que si $\det A(\xi) \equiv 0$, l'équation (1) n'admet une solution que pour certains f_n finis (en n).

2. Solution fondamentale. Nous appellerons *solution fondamentale du système* (1) une fonction matricielle G_n qui satisfait simultanément aux deux équations suivantes :

$$\sum_{k \in K} A_k G_{n+k} = \delta_n^0 E, \tag{4}$$

$$\sum_{k \in K} G_{n+k} A_k = \delta_n^0 E. \tag{4'}$$

LEMME. *Soit $Q(\xi_1, \ldots, \xi_t)$ un polynôme arbitraire en un nombre t quelconque d'arguments complexes, qui ne s'annule pas identiquement. On peut alors choisir les rayons r_j des circonférences $|\xi_j| = r_j$ de façon qu'on ait l'inégalité $Q(\xi_1, \ldots, \xi_t) \neq 0$ si $|\xi_1| = r_1, \ldots, |\xi_t| = r_t$.*

On mène la DÉMONSTRATION par récurrence sur t. Pour $t = 1$, le nombre de racines de $Q(\xi_1) = 0$ est fini et l'assertion du lemme est évidente. En supposant celle-ci démontrée pour $t = p$ établissons-la si $t = p+1$. Ordonnons le polynôme $Q(\xi_1, \ldots, \xi_{p+1})$ suivant les puissances de ξ_{p+1}:

$$Q(\xi_1, \ldots, \xi_{p+1}) = Q_0(\xi_1, \ldots, \xi_p)\xi_{p+1}^M + \ldots + Q_M(\xi_1, \ldots, \xi_p),$$

où M est un entier naturel et $Q_0(\xi_1, \ldots, \xi_p)$ ne s'annule pas identiquement. Choisissons $r_1, \ldots, r_p$ de sorte que $Q_0(\xi_1, \ldots, \xi_p) \neq 0$ pour $|\xi_1| = r_1, \ldots, |\xi_p| = r_p$, ce qui est possible par hypothèse de récurrence. Prenant un r_{p+1} suffisamment grand, on peut avoir $Q(\xi_1, \ldots, \xi_{p+1}) \neq 0$ pour $|\xi_j| = r_j, j = 1, \ldots, p+1$.

THÉORÈME 1. *Une matrice G_n définie par l'égalité*

$$G_n = \frac{1}{(2\pi i)^s} \oint \ldots \oint_{|\xi_j| = r_j} \ldots \oint \frac{A^{-1}(\xi)}{\xi_1^{n_1+1} \ldots \xi_s^{n_s+1}} d\xi_1 \ldots d\xi_s \tag{5}$$

est solution fondamentale.

Conformément au lemme, les r_j sont choisis de sorte que $\det A(\xi) \neq 0$ si $|\xi_j| = r_j$.

La DÉMONSTRATION est immédiate. Compte tenu des propriétés des résidus nous obtenons

$$\sum_{k \in K} A_k G_{n+k} = \frac{1}{(2\pi i)^s} \int \ldots \int \frac{A(\xi) A^{-1}(\xi)}{\xi_1^{n_1+1} \ldots \xi_s^{n_s+1}} d\xi_1 \ldots d\xi_s = \delta_n^0 E,$$

$$\sum_{k \in K} G_{n+k} A_k = \frac{1}{(2\pi i)^s} \int \ldots \int \frac{A^{-1}(\xi) A(\xi)}{\xi_1^{n_1+1} \ldots \xi_s^{n_s+1}} d\xi_1 \ldots d\xi_s = \delta_n^0 E.$$

3. Frontière d'un domaine discrétisé. Considérons l'équation (1) sur un ensemble borné

$$\mathbf{L}u \equiv \sum_{k \in K} A_k u_{n+k} = f_n, \quad n \in D_0, \tag{6}$$

avec D_0 un domaine discrétisé quelconque où f_n est défini. Le domaine de définition de la solution u_n est alors un ensemble D que parcourt le point $n+k$ si n et k parcourent indépendamment D_0 et K respectivement. Faisons correspondre à chaque $r \in D$ un sous-ensemble K_r de K, qui est formé de tous les $k \in K$ pour lesquels $r-k \notin D_0$. Appelons *frontière* Γ l'ensemble de tous les points $r \in D$ pour lesquels K_r n'est pas vide. Dans le cas de l'analogue aux différences très simple de l'équation de Poisson

$$\mathbf{L}u \equiv u_{n_1-1, n_2} + u_{n_1, n_2+1} + u_{n_1+1, n_2} + u_{n_1, n_2-1} - 4u_{n_1 n_2} = h^2 F_{n_1 n_2},$$

$$|n_1| < N, \quad |n_2| < N; \qquad Nh = 1,$$

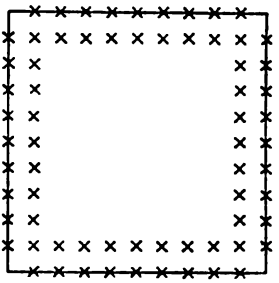


Fig. 52

l'ensemble D_0 se compose, par exemple, de tous les (n_1h, n_2h) tombés dans le carré $|x_1| \leqslant 1$, $|x_2| \leqslant 1$; l'ensemble K de cinq vecteurs $(1, 0)$, $(0, 1)$, $(-1, 0)$, $(0, -1)$, $(0, 0)$; l'ensemble D de tous les points entiers du carré $|n_1| \leqslant N$, $|n_2| \leqslant N$, moins les quatre points anguleux $|n_1| = |n_2| = N$; la frontière Γ de deux couches de points repérés par les petites croix (fig. 52).

4. Analogues aux différences des formules intégrales de Cauchy et du type de Cauchy.

LEMME. *Soit B_n une matrice fonctionnelle arbitraire définie sur tout le réseau entier pour laquelle la multiplication à droite par une matrice carrée d'ordre m a un sens. On a l'identité suivante :*

$$\sum_{n \in D_0} B_{-n} \sum_{k \in K} A_k u_{n+k} \equiv \sum_{n \in D} \Big(\sum_{k \in K} B_{-n+k} A_k\Big) u_n - \sum_{r \in \Gamma} \Big(\sum_{k \in K_r} B_{-r+k} A_k\Big) u_r. \tag{7}$$

DÉMONSTRATION. La fonction vectorielle u_n, $n \in D$, peut s'écrire

$$u_n = \sum_{t \in D} \delta_n^t u_n.$$

Les deux membres de (7) dépendent linéairement de u. Il suffit donc de vérifier l'identité (7) pour la fonction vectorielle

$$v_n \equiv v_n^t \equiv \delta_n^t u_n = \begin{cases} 0 & \text{si} \quad n \neq t, \\ u_t & \text{si} \quad n = t \end{cases}$$

pour tout $t \in D$ fixe :

$$\sum_{n \in D_0} B_{-n} \sum_{k \in K} A_k v_{n+k} \equiv \sum_{k \in K} \sum_{n \in D_0} B_{-n} A_k v_{n+k} \equiv$$

$$\equiv \sum_{k \in K} \delta_{D_0}^{t-k} B_{-t+k} A_k v_t \equiv \sum_{k \in K} B_{-t+k} A_k v_t - \sum_{k \in K_t} B_{-t+k} A_k v_t \equiv$$

$$\equiv \sum_{n \in D} \Big(\sum_{k \in K} B_{-n+k} A_k\Big) v_n - \sum_{r \in \Gamma} \Big(\sum_{k \in K_r} B_{-r+K} A_k\Big) v_r,$$

où

$$\delta_{D_0}^{t-k} = \begin{cases} 1 & \text{si} \quad t-k \in D_0. \\ 0 & \text{si} \quad t-k \notin D_0. \end{cases}$$

THÉORÈME 2. *Soient $\{u_n\}$, $n \in D$, une solution quelconque de l'équation* (6) *et G_n une solution fondamentale arbitraire. On a alors la formule*

$$\sum_{r \in \Gamma} \Big(\sum_{k \in K_r} G_{n-r+k} A_k\Big) u_r + \sum_{m \in D_0} G_{n-m} f_m = \begin{cases} u_n & \text{si} \quad n \in D, \\ 0 & \text{si} \quad n \notin D. \end{cases} \tag{8}$$

DÉMONSTRATION. Multiplions à gauche les deux membres de (6) par la matrice G_{t-n}, puis sommons en tous les $n \in D_0$. Moyennant l'identité (7) et l'égalité (4'), on obtient la formule (8).

CONSÉQUENCE. *Chaque solution $\{u_n\}$ de l'équation* (6) *est bien définie par ses valeurs sur Γ et on la cherche par la formule* (8) *connaissant ces valeurs.*

THÉORÈME 3. *Etant donné une fonction vectorielle quelconque $\{v_r\}$ de dimension m définie sur Γ et une solution fondamentale arbitraire G_n, la formule*

$$u_n = \sum_{r \in \Gamma} \Big(\sum_{k \in K_r} G_{n-r+k} A_k\Big) v_r + \sum_{m \in D_0} G_{n-m} f_m, \quad n \in D, \tag{9}$$

définit une solution de (6).

DÉMONSTRATION. Faisons agir l'opérateur L sur la fonction vectorielle $\{u_n\}$ définie par la formule (9) :

$$\mathbf{L}u_n \equiv \sum_{r\in\Gamma}\Big[\sum_{k\in K_r}(\mathbf{L}G_{n-r+k})A_k\Big]v_r + \sum_{m\in D_0}(\mathbf{L}G_{n-m})f_m, \quad n\in D_0. \tag{10}$$

Calculons le second membre de (10). Nous avons en vertu de (4)

$$\mathbf{L}G_{n-r+k} = \begin{cases} E & \text{si} \quad n = r-k, \\ 0 & \text{si} \quad n \neq r-k. \end{cases}$$

Par définition de l'ensemble K_r le point $n = r-k$ n'appartient pas à D_0 si bien que le premier terme de la somme de (10) est un vecteur nul. Le second terme est évidemment f_n de sorte que $\mathbf{L}u_n = f_n$, et le théorème est démontré.

L'égalité (8) est analogue à la formule intégrale de Cauchy pour les fonctions analytiques $\varphi(z)$ dans un domaine borné d de frontière γ :

$$\frac{1}{2\pi i}\oint_{\zeta\in\gamma}\frac{\varphi(\zeta)}{\zeta-z}\,d\zeta = \begin{cases} \varphi(z) & \text{si} \quad z\in d, \\ 0 & \text{si} \quad z\notin d\cup\gamma. \end{cases} \tag{11}$$

Ceci étant, le rôle de fonctions analytiques, de frontière du domaine et du noyau de Cauchy $\frac{1}{2\pi i}\frac{1}{\zeta-z}$ est joué respectivement par les solutions $\{u_n\}$ du problème (6), la frontière Γ du domaine discrétisé D et l'expression $\Big(\sum_{k\in K_r} G_{n-r+k}A_k\Big)$, l'ensemble K_r en lequel on somme mettant en évidence la structure de la frontière au voisinage du point $r\in\Gamma$.

La formule (9) évoque alors naturellement la formule intégrale du type de Cauchy. Quant à la formule (8), elle est analogue à celle de Green pour l'équation de Laplace.

Il existe néanmoins une différence essentielle entre (11) et (8) : la formule intégrale de Cauchy n'est valable que pour l'intérieur du domaine d et la formule aux différences (8) l'est partout sur D, *y compris aux points de la frontière* Γ.

On note une différence analogue en ce qui concerne (9) et la formule de Green.

5. Conditions aux limites intrinsèques.

THÉORÈME 4. *Soit G_n une solution fondamentale de l'équation* (1). *Pour qu'une fonction vectorielle $\{u_r\}$, $r\in\Gamma$, définie sur Γ soit prolongeable partout dans un domaine discrétisé borné D en une solution de l'équation* (6), *il faut et il suffit que tous les $n\in\Gamma$ vérifient les égalités*

$$\sum_{r\in\Gamma}\Big(\sum_{k\in K_r} G_{n-r+k}\Big)u_r + \sum_{m\in D_0} G_{n-m}f_m = u_n, \quad n\in\Gamma. \tag{12}$$

DÉMONSTRATION. Si $\{u_r\}$, $r\in\Gamma$, est prolongeable partout sur D en une solution $\{u_n\}$, $n\in D$, en appliquant à cette solution la formule (8), puis en considérant l'égalité obtenue pour $n\in\Gamma$ seuls, on s'assure que (12) est vrai. Inversement, si $\{u_r\}$, $r\in\Gamma$, satisfait à (12), on prend $v_r\equiv u_r$ et on construit une solution $\{u_n\}$, $n\in D$, d'après la formule (9). En vertu de (12), les valeurs frontières de cette solution $\{u_r\}$, $r\in\Gamma$, coïncident avec des valeurs données.

Le théorème que nous venons de démontrer nous autorise à appeler les égalités (12) *conditions aux limites intrinsèques :* ces conditions ne sont pas données de l'extérieur, ce sont les conséquences de l'équation aux différences même.

A condition d'interpréter (8) et (9) comme les analogues des formules intégrales de Cauchy et du type de Cauchy, on peut dire que les conditions aux limites intrinsèques sont analogues aux conditions classiques de Sokhotski-Plemelj sous lesquelles une fonction $\varphi(z)$ donnée sur la frontière γ d'un domaine d du plan complexe est prolongeable partout dans d en une fonction analytique.

La formule (8) peut s'interpréter comme la formule de Green aux différences adéquate au système (6), qui tient compte implicitement des « sauts de potentiel » sur la frontière Γ et conduit aux conditions (12).

6. Opérateur projection frontière. Les conditions aux limites intrinsèques se prêtent à une écriture différente de (12). Désignons par U_Γ l'espace vectoriel de toutes les fonctions vectorielles discrètes $u_\Gamma = \{u_r\}$, $r \in \Gamma$, et par U'_Γ le sous-espace formé de celles d'entre elles qui sont prolongeables partout dans D en des solutions $\{u_n\}$, $n \in D$, de l'équation homogène associée à (6).

Définissons l'application linéaire P, $u_\Gamma = Pv_\Gamma$, de l'espace U_Γ dans lui-même par la formule suivante :

$$u_n = \sum_{r \in \Gamma} \Big(\sum_{k \in K_r} G_{n-r+k} A_k \Big) v_r, \quad n \in \Gamma. \tag{13}$$

THÉORÈME 5. *P est l'opérateur projection de U_Γ sur U'_Γ.*

DÉMONSTRATION. En effet, quel que soit $v_\Gamma \in U_\Gamma$, l'élément $u_\Gamma = Pv_\Gamma$ appartient, en vertu du théorème 3, à U'_Γ. Si $u_\Gamma \in U'_\Gamma$, alors $Pu_\Gamma = u_\Gamma$ aux termes du théorème 2. Le théorème est démontré.

L'opérateur P défini par la formule (13) sera dit *opérateur projection frontière*. Moyennant cet opérateur les conditions aux limites intrinsèques (12) s'écrivent

$$u_\Gamma - Pu_\Gamma = 0 \tag{14}$$

dans le cas $f_n \equiv 0$. Soulignons que l'opérateur projection frontière dépend du choix de la solution fondamentale G_n.

7. Problème aux limites général. Etant donné la conséquence du théorème 2, on définit chaque solution de l'équation (6) connaissant ses valeurs sur la frontière Γ. C'est ce qui permet de déterminer le problème aux limites linéaire général pour (6) comme un problème de la forme

$$\left.\begin{aligned} \mathrm{L}u \equiv \sum_{k \in K} A_k u_{n+k} = f_n, \quad & n \in D_0, \\ lu_\Gamma = \varphi, \quad & \varphi \in \Phi, \end{aligned}\right\} \tag{15}$$

avec l un opérateur linéaire appliquant l'espace U_Γ sur un espace vectoriel Φ.

Les schémas aux différences naturels approchant le premier, le deuxième ou le troisième problème aux limites pour l'équation de Poisson s'écrivent par exemple facilement sous la forme (15).

L'appellation « problème aux limites général » est un peu conventionnelle, car il existe des problèmes aux limites aux différences d'une forme autre que (15). C'est le cas, par exemple, de schémas aux différences naturels pour les problèmes aux limites différentiels où l'équation différentielle est d'un ordre inférieur à celui des conditions aux limites différentielles.

8. Idée à la base de la méthode des conditions aux limites intrinsèques. Soit $f_n \equiv 0$ pour plus de simplicité. Il y a une relation étroite entre le problème aux limites aux différences

$$\mathrm{L}u = 0, \quad lu_\Gamma = \varphi \tag{16}$$

et le problème

$$u_\Gamma - Pu_\Gamma = 0, \quad lu_\Gamma = \varphi. \tag{17}$$

Notamment, les valeurs frontières $u_\Gamma = \{u_r\}$, $r \in \Gamma$, de chaque solution $\{u_n\}$, $n \in D$, du problème (16) vérifient, en vertu du théorème 2, les équations (17). Inversement, chaque solution $u_\Gamma = \{u_r\}$, $r \in \Gamma$, du problème (17) est prolongeable, selon le théorème 4 et la conséquence du théorème 2, de façon unique partout dans D en une solution du problème (16). L'idée fondamentale de la méthode des conditions aux limites intrinsèques consiste à passer du problème initial (16) au système d'équations (17) sur la frontière Γ. Les progrès ainsi obtenus sont fonction de deux circonstances suivantes : 1) un petit nombre (par rapport au problème (16)) d'inconnues du problème (17) ; 2) une forme spéciale du système (17) qui comprend comme partie intégrante l'opérateur projection frontière.

9. Stabilité des conditions aux limites intrinsèques. Il y a à craindre que les

conditions aux limites intrinsèques $u_\Gamma - Pu_\Gamma = 0$ ne soient « presque dégénérées », ce qui entraînerait un mauvais conditionnement du problème (17) quel que soit l'opérateur l, de sorte que le passage de (16) à (17) occasionne une perte de stabilité numérique.

Supposons que l'espace Φ est inclus dans U_Γ, introduisons dans ce dernier (donc dans $\Phi \subset U_\Gamma$) une norme $\|\cdot\|$ et démontrons un théorème aux termes duquel on passe du problème (16) au problème (17) sans qu'il y ait perte de stabilité numérique.

THÉORÈME 6. *Soit u_Γ solution du problème* (17) *pour tout $\varphi \in \Phi$, qui s'évalue par*

$$\|u_\Gamma\| \leqslant c\,\|\varphi\|, \tag{18}$$

c étant indépendant de φ. Soit ensuite v_Γ un élément quelconque de U_Γ. Notons

$$v_\Gamma - Pv_\Gamma = \tilde{\psi}, \qquad lv_\Gamma = \tilde{\varphi}. \tag{19}$$

On a alors l'estimation

$$\|v_\Gamma\| \leqslant c\,(\|\tilde{\varphi}\| + \|l\|\,\|\tilde{\psi}\|) + \|\tilde{\psi}\|. \tag{20}$$

A condition de prendre (19) pour équations définissant v_Γ, l'estimation (20) signifie que la sensibilité de la solution de (19) aux perturbations $\tilde{\psi}$ sur le second membre des conditions aux limites intrinsèques se caractérise par la constante c de l'estimation (18), i. e. par sa sensibilité aux perturbations sur le second membre de la condition aux limites donnée $lu_\Gamma = \varphi$.

DÉMONSTRATION. Notons $z_\Gamma \equiv v_\Gamma - \tilde{\psi} = Pv_\Gamma$. En vertu du théorème 5, $z_\Gamma = Pv_\Gamma \in U'_\Gamma$. Donc

$$z_\Gamma - Pz_\Gamma = 0, \qquad lz_\Gamma = l(u_\Gamma - \tilde{\psi}) = \tilde{\varphi} - l\tilde{\psi},$$

i.e. z_Γ vérifie un système de la forme (17) et, selon (18), l'estimation

$$\|z_\Gamma\| \leqslant c\,\|\tilde{\varphi} - l\tilde{\psi}\| \leqslant c(\|\tilde{\varphi}\| + \|l\|\,\|\tilde{\psi}\|).$$

Compte tenu de l'identité $z_\Gamma = v_\Gamma - \tilde{\psi}$, il en résulte (20).

10. Une idée supplémentaire. Exposons une idée à utiliser pour résoudre sur machine les problèmes aux limites relatifs à des équations aux dérivées partielles pour le problème suivant.

Soit $u(x, y)$ une fonction définie dans un domaine d de frontière γ suffisamment régulière en tant que solution du problème de Dirichlet

$$\left.\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= 0, \qquad (x, y) \in d, \\ u|_\gamma &= a(s). \end{aligned}\right\}$$

On demande de trouver la dérivée $\left.\dfrac{\partial u}{\partial n}\right|_\gamma = b(s)$ dans la direction de la normale intérieure Ce problème intervient si, connaissant la température $u|_\gamma = a(s)$ sur γ, on veut trouver le flux thermique stationnaire à travers cette frontière.

On entend par s la longueur d'arc suivant γ et on estime, pour fixer les idées, que la longueur totale de γ vaut 2π. Cherchons $\left.\dfrac{\partial u}{\partial n}\right|_\gamma = b(s)$ de façon approchée comme somme partielle

$$b(s) = \sum_{j=0}^{k} (\alpha_j \cos js + \beta_j \sin js)$$

de sa série de Fourier. Les coefficients α_j et β_j se déterminent par la méthode des conditions aux limites intrinsèques. Donnons-nous $h > 0$, construisons le réseau $(x_{n_1}, y_{n_2}) = (n_1 h, n_2 h)$, l'équation aux différences

$$u_{n_1+1,\, n_2} + u_{n_1-1,\, n_2} + u_{n_1,\, n_2+1} + u_{n_1,\, n_2-1} - 4u_{n_1,\, n_2} = 0$$

et rapportons à $D_0 = D_0^h$ tous les points du réseau qui appartiennent, de même que les quatre points voisins, à $d \cup \gamma$. On définit ainsi un domaine discrétisé $D = D^h$, sa frontière $\Gamma = \Gamma^h$ et les conditions aux limites intrinsèques $u_\Gamma - Pu_\Gamma = 0$.

L'idée consiste à prolonger, connaissant $u|_\gamma = a(s)$ et $\left.\frac{\partial u}{\partial n}\right|_\gamma = b(s)$ écrit sous forme de série à coefficients indéterminés, la solution moyennant la formule de Taylor de la frontière γ dans la bande limitrophe contenant la frontière Γ^h du domaine discrétisé, puis à choisir les coefficients indéterminés $\alpha_j = \alpha_j^h$, $\beta_j = \beta_j^h$ à partir de la condition de minimisation du résidu apparu par substitution de la fonction $u(x, y)$ ainsi prolongée dans les conditions aux limites intrinsèques.

Détaillons l'algorithme. Connaissant $u|_\gamma = a(s)$ et $\left.\frac{\partial u}{\partial n}\right|_\gamma = b(s)$ on trouve, en vertu de l'équation de Laplace, les fonctions $\left.\frac{\partial^2 u}{\partial n^2}\right|_\gamma$ et $\left.\frac{\partial^3 u}{\partial n^3}\right|_\gamma$, combinaisons linéaires des fonctions $a(s)$, $a'(s)$, $a''(s)$, $b(s)$, $b'(s)$, $b''(s)$ à coefficients bornés dépendant de la seule courbure de la frontière γ.

Dans les expressions de $\left.\frac{\partial u}{\partial n}\right|_\gamma$, $\left.\frac{\partial^2 u}{\partial n^2}\right|_\gamma$ et $\left.\frac{\partial^3 u}{\partial n^3}\right|_\gamma$ laissons uniquement les termes précédés de coefficients α_j et β_j, $j \leqslant k$; le nombre k sera précisé plus loin. Soient $q \in \Gamma$ et $q' \in \Gamma$ le point le plus proche de q sur γ.

Définissons approximativement u_q par la formule

$$u_q = a(s)\,|_{s=s_{q'}} + qq' \left.\frac{\partial u}{\partial n}\right|_{s=s_{q'}} + \frac{(qq')^2}{2!} \left.\frac{\partial^2 u}{\partial n^2}\right|_{s=s_{q'}} + \frac{(qq')^3}{3!} \left.\frac{\partial^3 u}{\partial n^3}\right|_{s=s_{q'}}$$

et portons ces expressions pour $q \in \Gamma$ dans le premier membre de la condition intrinsèque $u_\Gamma - Pu_\Gamma = 0$. On voit surgir le résidu $\psi_\Gamma = u_\Gamma - Pu_\Gamma$, fonction du choix des coefficients indéterminés α_j, β_j, $j \leqslant k$.

THÉORÈME 7. *Supposons que la solution $u(x, y)$ du problème de Dirichlet possède des dérivées bornées jusqu'à l'ordre $r \geqslant 6$ dans le domaine d jusqu'à sur la frontière γ et qu'en construisant les conditions aux limites intrinsèques on prend pour G_n une solution fondamentale qui croît pour $n_1^2 + n_2^2 \to \infty$ comme logarithme $\int \sqrt{n_1^2 + n_2^2}$. Cette solution fondamentale a été obtenue et étudiée par S. Sobolev* (Doklady Akad. Nauk S.S.S.R., **87**, 3 (1952), pp. 341-344).

Posons $k = h^{-\frac{3}{r-2}}$. *Si* $\alpha_j = \alpha_j^h$, $\beta_j = \beta_j^h$, $j \leqslant k$, *sont définis par la condition*

$$\max_{q \in \Gamma} |\psi_q| = \min$$

ou par

$$\sum_{q \in \Gamma} |\psi_q|^2 = \min,$$

on a les égalités respectives suivantes :

$$\max_s \left| b(s) - \sum_{j=0}^{k} (\alpha_j \cos js + \beta_j \sin js) \right| = O(h),$$

$$\max_s \left| b(s) - \sum_{j=0}^{k} (\alpha_j \cos js + \beta_j \sin js) \right| = O(\sqrt{h}).$$

La démonstration étant volumineuse, nous la laissons de côté.

Notons que plus r est grand, moins k est élevé et plus le calcul de α_j, β_j, $j \leqslant k$, est facile.

11. Comparaison de la méthode des conditions aux limites intrinsèques avec la méthode des équations intégrales singulières. Au début de l'Annexe nous avons noté une analogie entre la méthode des conditions aux limites intrinsèques et celle des équations intégrales singulières. Comparons ces méthodes afin de préciser la ressemblance et les différences essentielles.

Illustrons, pour commencer, l'idée de la deuxième méthode pour les problèmes aux limites différentiels sur l'exemple du problème

$$\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}-\mu u = 0, \qquad x = (x_1, x_2) \in \Omega, \tag{21}$$

$$a_0 u_0 + a_1 u_1 = \varphi(x), \qquad x = (x_1, x_2) \in \omega, \tag{22}$$

où $\mu = \text{const} > 0$, Ω est un domaine borné et ω sa frontière. La condition aux limites (22) relie la solution $u = u_0(x)$ sur la frontière du domaine et sa dérivée $\partial u/\partial \nu = u_1(x)$ dans la direction de la normale intérieure. Les coefficients a_0 et a_1 sont des opérateurs donnés.

Ecrivons la formule classique de Green pour l'équation (21) :

$$u(x) = \int\limits_{y \in \omega} \left[g(x-y)\frac{\partial u}{\partial \nu} - u\frac{\partial g}{\partial \nu}\right] d\sigma_y, \tag{23}$$

où $g(x)$ est solution fondamentale de (21) tendant vers zéro à l'infini. Faisons tendre x vers ω. Compte tenu des propriétés des potentiels de simple et de double couche nous obtenons sur la frontière ω une relation de la forme

$$u_0 = b_0 u_0 + b_1 u_1 \tag{24}$$

entre la solution $u(x)$ et sa dérivée normale $\partial u/\partial \nu = u_1(x)$ sur ω ; b_0 et b_1 sont des opérateurs intégraux connus. Le passage du problème (21), (22) au système équivalent (22), (24) par rapport aux fonctions $u_0(x)$, $u_1(x)$ définies sur la frontière ω constitue justement l'essence de la méthode des équations intégrales singulières.

Adressons-nous maintenant à la méthode des conditions aux limites intrinsèques appliquée au problème aux limites général suivant pour l'analogue aux différences de l'équation (21) dans un domaine discret carré

$$u_{n_1-1,\, n_2} + u_{n_1,\, n_2+1} + u_{n_1+1,\, n_2} + u_{n_1,\, n_2-1} - (4+\mu) u_{n_1,\, n_2} = 0, \tag{25}$$

$$-N < n_1, \quad n_2 < N,$$

$$l u_\Gamma = \varphi. \tag{26}$$

Mettons les conditions aux limites intrinsèques $u_\Gamma - P u_\Gamma = 0$ sous une forme commode pour l'exposé qui suit.

On vérifie aisément que la formule (9) se récrit alors

$$u_n \equiv \sum_{r \in Q_0} [G_{n-r}(\Delta_r u_r) - u_r(\Delta_r G_{n-r})] + \sum_{r \in Q_0} u_r \delta_n^r, \qquad n \in D, \tag{27}$$

avec Q_0 l'ensemble des points de Γ situés sur les côtés du carré $|n_1| = N$, $|n_2| = N$, i.e. sur la face extérieure de la frontière bilatérale Γ du réseau carré (fig. 52), et Δ_r l'analogue aux différences de la dérivée dans la direction de la normale intérieure.

La formule (27) serait un analogue parfait de la formule classique de Green (23) n'était-ce le « terme singulier » $\sum \delta_n^r u_r$ de son second membre. Si ce terme était absent, l'égalité (27) ne serait vérifiée que pour $n \in D_0$ et elle ne donnerait pas lieu aux conditions aux limites intrinsèques $u_\Gamma - P u_\Gamma = 0$. Celles-ci s'obtiennent de (27) si n parcourt seulement les points de la frontière Γ de D et non pas ce domaine tout entier, et on les écrit à l'aide de deux systèmes d'égalités

$$u_n = \sum_{r \in Q_0} [G_{n-r}(\Delta_r u_r) - u_r(\Delta_r G_{n-r})] + u_n, \qquad n \in Q_0, \tag{28}$$

$$u_n = \sum_{r \in Q_0} [G_{n-r}(\Delta_r u_r) - u_r(\Delta_r G_{n-r})], \qquad n \in \Gamma \backslash Q_0, \tag{29}$$

correspondant respectivement aux points $n \in Q_0$ de la face extérieure de Γ et aux points $n \in \Gamma \backslash Q_0$ de sa face intérieure. On prendra pour G_n dans les égalités (28) et (29) une solution fondamentale bornée.

On montre que les conditions aux limites intrinsèques $u_\Gamma - Pu_\Gamma = 0$, i.e. le système d'équations (28), (29) est algébriquement équivalent à chacun des sous-systèmes (28) et (29) pris isolément.

Le sous-système (28) est analogue à la relation intégrale (24) si bien que l'analogue aux différences du problème (22), (24) est le problème (27), (28) et non pas

$$lu_\Gamma = \varphi, \quad u_\Gamma - Pu_\Gamma = 0,$$

problème décrit par les égalités (27) à (29) qui nous sert à illustrer la méthode des conditions aux limites intrinsèques.

Il y a une différence évidente entre les conditions $u_\Gamma - Pu_\Gamma = 0$, i.e. le système (28), (29), et le seul sous-système (28). Les premières contiennent les égalités surabondantes (29). En ce sens, les conditions aux limites aux différences intrinsèques $u_\Gamma - Pu_\Gamma = 0$ s'apparentent aux conditions de Sokhotski-Plemelj pour les fonctions analytiques plutôt qu'à la relation intégrale (24). Les conditions de Sokhotski-Plemelj sont deux relations réelles relatives à deux fonctions réelles, mais elles ne sont pas indépendantes, et la variété de couples de fonctions les vérifiant dépend d'une fonction réelle arbitraire.

Notons que le fait de comprendre l'opérateur projection frontière avantage heureusement les conditions aux limites intrinsèques $u_\Gamma - Pu_\Gamma = 0$ par rapport au sous-système équivalent (28). C'est la raison pour laquelle le problème $lu_\Gamma = \varphi, u_\Gamma - Pu_\Gamma = \psi$ est stable au sens du théorème 6 par rapport à la perturbation sur le second membre ψ. En supprimant dans le cas général certaines équations du système $u_\Gamma - Pu_\Gamma = 0$ on peut aboutir à un sous-système qui équivaut du point de vue algébrique au système initial, mais qui n'est plus stable.

On montre que dans notre exemple (25), (26), on préférera, pour remplacer $u_\Gamma - Pu_\Gamma = 0$, au sous-système (28), analogue à la relation intégrale (24), le sous-système (29) qui est stable et se compose, à la différence de (28), d'équations indépendantes : son rang est égal au nombre d'équations constituantes.

Ainsi, dans l'exemple considéré, l'analogie entre la méthode des conditions aux limites intrinsèques et celle des équations intégrales singulières n'est pas parfaite. Cela est vrai à plus forte raison en ce qui concerne la méthode classique des équations intégrales dans laquelle on recherche non pas la solution du problème initial (21), (22) sur la frontière, mais une densité auxiliaire de potentiel de simple ou de double couche.

COMMENTAIRES

Ch. 1. §§ 1, 2. On peut faire connaissance avec la théorie générale des équations aux différences linéaires en consultant par exemple [18, chap. V].

Ch. 1. § 3, n° 3. Le théorème sur la disposition des racines a été indiqué par S. Godounov.

Ch. 2, § 5. Les auteurs ont pris connaissance de la méthode du balayage et de sa justification pour une classe de problèmes aux limites aux différences lorsqu'ils ont lu en 1953 le manuscrit de l'article de I. Guelfand et O. Lokoutsievski « Méthode du balayage de résolution d'équations aux différences » (voir p. ex. [3] ou [4]). Il existe des variantes du balayage destinées à calculer les solutions de problèmes aux limites aux différences autres que ceux abordés dans le présent ouvrage. Pour les résultats et la bibliographie voir p. ex. [6], [12], [15].

Ch. 3. C'est N. Bakhvalov qui a proposé d'utiliser, pour légitimer le balayage, la propriété de bon conditionnement d'un problème aux limites aux différences. Son idée a été partiellement réalisée dans [4], puis par V. Ognéva, Zh. Vychisl. Mat. i Mat. Fiz. 7, n° 4 (1967) qui s'est servie la première des systèmes tronqués. Un exposé modifié du dernier ouvrage est contenu dans [17].

En justifiant le bon conditionnement d'un problème aux limites aux différences (§ 6), les auteurs utilisent la thèse de l'étudiant de l'Université de Novosibirsk Baguisbaev (en particulier, l'exemple montrant qu'on ne peut ignorer la condition de régularité des coefficients). Le travail de Baguisbaev a été fait sous la direction de S. Godounov.

Ch. 6. §§ 19, 20. Celui qui veut faire une plus ample connaissance avec les méthodes numériques de résolution des équations différentielles ordinaires peut s'adresser aux ouvrages [3], [15] et à la bibliographie qu'ils contiennent.

Les schémas aux différences pour certaines classes importantes d'équations différentielles à coefficients discontinus ont été construits dans la théorie des schémas aux différences homogènes de A. Tikhonov et A. Samarski et exposés dans l'un des chapitres de [12].

Les schémas aux différences bilatéraux relatifs à des problèmes à deux et à n points et à des problèmes linéaires plus généraux pour les équations différentielles ordinaires d'ordre pair quelconque ont été proposés par E. Volkov, Trudy Mat. Inst. Steklov, **128** (1972). Le même auteur a obtenu des estimations de l'erreur sur une solution aux différences d'un problème à 2 points pour une équation du second ordre, estimations qui s'expriment explicitement en fonction des coefficients de l'équation (Trudy Mat. Inst. Steklov, **112** (1971)).

Ch. 7, § 21. On doit la première description du concept de la stabilité des schémas aux différences par rapport aux erreurs d'arrondi commises sur les données initiales à J. von Neumann et R. D. Richtmyer (*A Method for the Numerical Calculations of hydrodynamical shocks*. Jour. of applied physics, vol. 21, N°3, 1950). La première classification des définitions de la stabilité et de l'approximation dans laquelle ces propriétés entraînent la convergence a été proposée par V. Riabenki (Doklady Akad. Nauk S.S.S.R.

86, N°6 (1952)) dans le cas d'analogues aux différences du problème de Cauchy pour les systèmes d'équations aux dérivées partielles.

La classification des définitions fondamentales et le théorème aux termes duquel la convergence découle de l'approximation et de la stabilité sont proches de ceux de A. Philippov (Doklady Akad. Nauk S.S.S.R. **100**, N°6 (1955)). Voir également [11] ou [4]. La différence principale est que les auteurs donnent une définition plus universelle de l'approximation.

Il existe d'autres systèmes naturels de définitions fondamentales dans lesquels l'approximation et la stabilité garantissent la convergence. Celui de D. Lax (1956) en est le plus connu (voir p. ex. [19]). Sa théorie considère les schémas aux différences pour des problèmes instationnaires sous l'hypothèse que ceux-là agissent non pas dans l'espace de fonctions discrètes mais dans le même espace fonctionnel que l'équation différentielle. On démontre sous cette hypothèse (supplémentaire) que pour un schéma aux différences approximant il y a à la fois stabilité et convergence. Ce théorème d'équivalence particularise la construction plus générale de L. Kantorovitch (Uspehi Mat. Nauk **3**, fasc. 6 (1948)).

Ces dernières années A. Samarski a proposé et développé de commun avec A. Gouline une théorie de la stabilité applicable à une très vaste classe de schémas aux différences ([12], [13] et § 41 du présent ouvrage).

Pour les résultats récents, la bibliographie et l'exposé sommaire des travaux sur la stabilité des schémas aux différences voir [4], [7], [10] à [14], [19].

Il faut dire que le travail de R. Courant, K. Friedrichs et H. Lewy (*Über die partiellen Differentialgleichungen der mathematischen Physik*, Math. Annalen, V. 100, 1928) et de nombreux autres ouvrages où la méthode des différences finies sert à démontrer l'existence de solutions d'équations différentielles établissent des inégalités qui s'interprètent aujourd'hui comme la stabilité pour telle ou telle norme. La notion de stabilité est cependant consécutive à l'utilisation de schémas aux différences pour le calcul approché des solutions supposées existantes. Aussi la stabilité est-elle en général étudiée pour des normes plus faibles qu'il n'est nécessaire pour démontrer l'existence.

Le premier à recourir à la méthode des différences finies pour prouver l'existence de solutions d'équations aux dérivées partielles a été L. Lusternik (1924) (voir Uspehi Mat. Nauk **8** (1940)) qui considérait l'équation de Laplace.

Notons que E. Volkov (Trudy Mat. Inst. Steklov **117** (1972)) a évalué l'erreur sur la solution aux différences de l'équation de Poisson en l'absence de l'approximation d'ordre deux de l'opérateur de Laplace parmi les niveaux dont le nombre croît indéfiniment avec le fractionnement du pas. Cette estimation est en même temps plus forte qu'une approximation uniforme d'ordre deux vu qu'elle établit une décroissance complémentaire des erreurs au voisinage de la frontière du domaine.

Ch. 7, § 22, n°3. Le procédé exposé de construction de schémas aux différences a été avancé par P.L.I. Brian, A.I.Ch. E.J. **7** (1961); J. Douglas, Num. Math. **4** (1962); J. Douglas, Trans. Amer. Soc. **89** (1958) ; S. Godounov, Méthodes aux différences de résolution des équations de la mécanique des fluides, Novosibirsk, 1962 (rotaprint) (éd. russe). La version bidimensionnelle du schéma prédicteur-correcteur de Lax-Wendroff [19] pour les problèmes de mécanique des fluides a été proposée par L. Tchoudov (voir l'aperçu de G. Rosliakov et G. Télénine dans le recueil *Méthodes numériques en mécanique des fluides*, M., éd. Univ. Moscou, fasc. 2, 1963). L'idée de la méthode de Runge-Kutta a été utilisée par V. Roussanov (préprint de l'Institut de mathématiques appliquées de l'Acad. des Sci. de l'U.R.S.S., 1967) afin de construire un schéma aux différences exact à l'ordre 3 pour le calcul en mécanique des fluides.

L. Tchoudov (article dans le recueil *Certaines applications de la méthode des réseaux en mécanique des fluides*, fasc. 1, « Courants dans la couche limite », éd. Univ. Moscou, 1971) a construit pour les équations du type parabolique un schéma aux différences du type Runge-Kutta exact à l'ordre 2 qui possède de bonnes propriétés de lissage. Les schémas de prédiction-correction s'appliquent souvent dans le calcul en mécanique des fluides. Voir p. ex. [1]. Il existe des méthodes variationnelles et autres de construction de schémas aux différences (voir [7], [10] à [15], [19], [20]).

Ch. 8, § 25, n°5. Il paraît que la possibilité d'utiliser les approximations différentielles pour étudier les équations aux différences a été mise en évidence aux années

1950 par A. Joukov (communication à un séminaire de l'Institut de mathématiques appliquées) à qui on doit également l'exemple considéré. La théorie des approximations différentielles qui étudie les propriétés asymptotiques et de groupe de certaines classes intéressantes d'équations aux différences a été construite par N. Yanenko et Y. Chokine, Sib. Mat. Zh. **10**, N° 5 (1969) ; Méthodes numériques en mécanique des milieux continus **2**, N° 2 (1971). Ces questions sont également traitées par N. Kouznétsov, Doklady Akad. Nauk S.S.S.R. **200**, N° 5 (1971) ; Doklady Akad. Nauk S.S.S.R. **204**, N° 2 (1972) ; Zh. Vychisl. Math. i Mat. Fiz. **12**, N° 2 (1972).

Ch. 8, § 26, n°1. L'idée de fixer les coefficients aux points intérieurs a été avancée dans l'article cité (commentaire du § 21) de von Neumann et Richtmyer.

Ch. 8, § 26, n°2. Le critère de Babenko et Guelfand a été proposé dans leur rapport commun avec O. Lokoutsievski à la conférence de l'analyse fonctionnelle (Moscou, 1956). Voir également [2] et les commentaires du ch. 13.

Ch. 8, § 27. Il existe un algorithme de calcul des coefficients de la série finie de Fourier, dit transformation de Fourier rapide, qui est très économique en ce qui concerne le nombre d'opérations arithmétiques. Voir p. ex. [15] ou [7].

La méthode de Fourier a été utilisée par O. Ladygenskaïa pour étudier des équations aux différences (Rapport au Séminaire Petrovsky à l'Université de Moscou (1949) et Doklady Akad. Nauk S.S.S.R. **88**, N° 4 (1953)).

Ch. 8, § 28, n°2. O. Ladygenskaïa a été la première à construire un schéma aux différences implicite pour les systèmes hyperboliques au sens de Petrovsky (Doklady Akad. Nauk S.S.S.R. **88**, N° 4 (1953)).

Ch. 9. Voir [10] et sa bibliographie ; on trouve dans les recueils et les revues de nouveaux ouvrages consacrés aux méthodes numériques en mécanique des milieux continus.

Ch. 10. Le schéma de directions alternées (12) du § 32 a été construit par D. Peaceman et H. Rachford en 1956 (voir p. ex. [16] ou [20]) ; le schéma de décomposition (7) du § 31 a été proposé par N. Yanenko, Doklady Akad. Nauk S.S.S.R., **125**, N° 6 (1959). On dispose aujourd'hui de tels schémas pour nombre de problèmes fondamentaux de la physique mathématique. Voir p. ex. [7], [12], [14], [16], [20], la monographie de E. Diakonov, *Méthodes aux différences de résolution des problèmes aux limites*, 1re partie (1971), 2ième partie (1972), éd. Univ. Moscou, et la bibliographie qu'elle contient.

En ce qui concerne la méthode de grosses particules de Bélotserkovski et Davydov et ses applications, consulter, en plus du travail de ces auteurs cité dans le § 33, le recueil d'articles, *Problèmes de mathématiques appliquées et de mécanique*, M., « Naouka » (1971), Zh. Vychisl. Mat. i Mat. Fiz. **13**, N° 1 (1973).

Ch. 11, § 34. Pour calculer la solution de l'équation de Poisson aux différences dans un rectangle le procédé le plus économique est la transformation de Fourier rapide (voir commentaire du § 27). Les schémas aux différences pour les équations de Laplace et de Poisson dans des domaines curvilignes ont intéressé, depuis Lusternik (1924), de nombreux savants. Voir p. ex. [8], [12], [16] et leur bibliographie.

Pour plusieurs schémas approximant les problèmes de Dirichlet, de Neumann et le problème aux limites mixte relatif aux équations de Laplace et de Poisson sur un rectangle, un parallélépipède rectangle et certains triangles on a obtenu des estimations d'erreur s'exprimant directement par les données initiales (E. Volkov, Trudy Mat. Inst. Steklov, **74** (1966), **105** (1969), I. Soultanova, Zh. Vychisl. Mat. i Mat. Fiz, **11**, N° 5 (1971) et la bibliographie de cet article). E. Volkov a également établi (Trudy Mat. Inst. Steklov, **128** (1972)) que si l'opérateur aux différences satisfait aux nœuds frontières à la condition d'être adéquat au laplacien aux différences à cinq points standard, l'équation aux différences de Poisson prolongée du réseau à un domaine fermé de frontière curviligne approche, pour des données suffisamment régulières, à l'ordre 2 par rapport au pas la solution cherchée avec ses dérivées jusqu'à l'ordre n inclus, $n \geqslant 0$ quelconque.

Selon E. Volkov (Trudy Mat. Inst. Steklov, **96** (1968)), la méthode des réseaux carrés sur un domaine d'angle $\alpha\pi$, $1/2 < \alpha \leqslant 2$, $\alpha \neq 1$, garantit la convergence uniforme de la solution aux différences de l'équation de Laplace à l'ordre $1/\alpha$ au plus, les valeurs frontières pouvant être aussi régulières qu'on le veut. On trouve dans le même article une méthode

des réseaux carrés et polaires composés pour l'équation de Laplace sur des domaines finis et infinis de frontière régulière par morceaux, qui garantit, dans des conditions naturelles sur les valeurs frontières, une convergence uniforme de l'ordre de $O(h^2)$ de la solution aux différences sur un réseau dont les nœuds sont au nombre de $O(h^{-2} \ln h^{-1})$.

Ch. 11, § 35. L'idée de considérer les solutions des problèmes stationnaires comme limite de celles des problèmes instationnaires avec l'augmentation du temps a été utilisée pour la première fois dans les années 1930 par A. Tikhonov.

Un schéma aux différences de stationnarisation pour le calcul de l'écoulement supersonique stationnaire des corps par un gaz a été proposé par S. Godounov, A. Zabrodine et G. Prokopov, Zh. Vychisl. Mat. i Mat. Fiz. **1**, N° 6 (1961). Il est à noter qu'en justifiant la stabilité de ce schéma K. Bagrinovski et S. Godounov (Doklady Akad. Nauk S.S.S.R., **115**, N° 3 (1957)) utilisent l'éclatement des opérateurs aux différences. De nombreux auteurs se sont occupés du calcul de problèmes stationnaires par stationnarisation.

L. Lusternik a indiqué une des premières méthodes efficaces d'accélération de la convergence dans la résolution de l'équation de Poisson aux différences (Trudy Mat. Inst. Steklov, **20** (1947)).

Ch. 11, § 36. A commencer par les travaux de A. Abramov, M. Gavourine, Flanders et Shortley datant de 1950, on utilise dans divers problèmes les polynômes de Tchebycheff pour choisir un ensemble optimal de paramètres d'itération.

Pour les résultats récents, la bibliographie et divers aperçus des méthodes itératives de résolution de problèmes aux limites elliptiques aux différences voir entre autres [8], [12], [16], [20] ; les monographies de E. Diakonov, *Méthodes itératives de résolution d'analogues aux différences des problèmes aux limites pour les équations du type elliptique*, Kiev, 1970 (rotaprint), de G. Martchouk et Y. Kouznétsov, *Méthodes itératives et fonctionnelles quadratiques*, Novisibirsk, « Naouka », Section sibérienne, 1972 (rotaprint) ; l'article de synthèse de R. Fédorenko, Uspehi Mat. Nauk **28**, N° 2 (1973).

Ch. 12, § 40. Les solutions stationnaires sont souvent utilisées pour éclaircir le caractère de la convergence au voisinage des frontières. Voir p. ex. S. Godounov, Mat. sb. **47** (89), 3 (1957).

Ch. 12, § 41, n° 4. Les auteurs ont utilisé le n° 4 du § 6 de [11] dû à A. Philippov.

Ch. 12, § 41, n° 5. Le choix du produit scalaire $(u, v)_{B_k}$ par la formule (21) a été visiblement proposé par H. Mignot en 1953 pour un cas particulier de l'analogue aux différences de l'équation de la chaleur à coefficients variables et, sous une forme plus générale, dans le § 15 de [11] qui contient également un exposé modifié du travail mentionné de Mignot.

Ch. 12, § 41, n° 6. Le premier critère de stabilité de Samarski de ce numéro s'obtient du théorème 5, n° 6, § 1, ch. VI de [12] si au lieu d'un espace de Hilbert on considère l'espace euclidien et si on pose $\varrho = 1$. Voir également n° 7, § 1, ch. VI de [12].

Ch. 13, § 42. La notion de spectre d'une famille d'opérateurs aux différences a été introduite dans [4] où elle a permis en particulier de justifier le critère de stabilité des problèmes instationnaires sur un segment dû à Babenko et Guelfand. On y démontre de même qu'une condition nécessaire de stabilité est que le spectre de la famille soit dans le cercle unité.

Le théorème 2 a été obtenu par V. Riabenki, Doklady Akad. Nauk S.S.S.R., **185**, N° 2 (1969).

Ch. 13, § 44. La notion de noyau du spectre d'une famille d'opérateurs a été introduite par V. Riabenki, Doklady Akad. Nauk S.S.S.R. **185**, N° 2 (1969). L'auteur y a également formulé les théorèmes 1 à 4.

Le théorème de A. Sokolov pour les coefficients scalaires A_k, B_k a été publié dans Doklady Akad. Nauk S.S.S.R. **208**, N° 2 (1973) et la démonstration dans le cas général de coefficients matriciels a été communiquée par le même auteur après la parution de l'article indiqué.

Ch. 13, § 45. Le paragraphe reproduit l'article de V. Riabenki, Doklady Akad. Nauk S.S.S.R. **193**, N° 3 (1970).

Annexe. La méthode des conditions aux limites intrinsèques (MCLI) a été proposée par V. Riabenki, Thèse de doctorat, Institut de mathématiques appliquées de l'Académie des Sciences de l'U.R.S.S. (1969). Les n°s 1 à 9 et 11 exposent partielle-

ment l'article de V. Riabenki (Uspehi Mat. Nauk **26**, N° 3 (1971)). L'article en question fournit de plus certaines applications de la méthode à l'étude et au calcul des solutions de problèmes aux limites aux différences dans des domaines simples et composés.

Le contenu du n° 10 reproduit le rapport de V. Riabenki à la conférence célébrant les soixante-quinze ans de l'académicien Petrovsky (Université de Moscou, janvier 1976).

Annexe, n° 2. A. Béliankov (Math. Zametki **18**, N° 5 (1975)) a démontré l'existence d'une solution fondamentale qui croît pour $\|n\|^2 = n_1^2 + \ldots + n_s^2 \to \infty$ au plus comme une puissance de $\|n\|$.

Le même auteur a construit une solution fondamentale dite cyclique permettant de former les conditions aux limites intrinsèques et admettant une construction efficace à l'aide de la transformation de Fourier rapide.

A. Zabrodine et V. Ognéva (préprint de l'Institut de mathématiques appliquées de l'Académie des Sciences de l'U.R.S.S. (1973)) ont utilisé une modification à eux de MCLI pour calculer un problème non linéaire de la chaleur sur les graphes.

BIBLIOGRAPHIE

[1] АЛАЛЫКИН Г. Б., ГОДУНОВ С. К., КИРЕЕВА И. Л., ПЛИНЕР Л. А., Решение одномерных задач газовой динамики в подвижных сетках, М., « Наука », 1970.

[2] БАБЕНКО К. И., ВОСКРЕСЕНСКИЙ Г. П., ЛЮБИМОВ А. Н., РУСАНОВ В. В., Пространственное обтекание гладких тел идеальным газом, М., « Наука », 1964.

[3] БЕРЕЗИН И. С., ЖИДКОВ Н. П., Методы вычислений, т. 2, М., Физматгиз, 1959.

[4] ГОДУНОВ С. К., РЯБЕНЬКИЙ В. С., Введение в теорию разностных схем, М., Физматгиз, 1962.

[5] ДЬЯЧЕНКО В. Ф., Основные понятия вычислительной математики, М., « Наука », 1972.

[6] ЛЮБИМОВ А. И., РУСАНОВ В. В., Течение газа около тупых тел, ч. 1, М., 1970.

[7] МАРЧУК Г. И., Методы вычислительной математики, Новосибирск, « Наука », 1973.

[8] МАРЧУК Г. И., ЛЕБЕДЕВ В. И., Численные методы в теории переноса нейтронов, М., Атомиздат, 1971.

[9] ПЕТРОВСКИЙ И. Г., Лекции об уравнениях с частными производными, М., Физматгиз, 1961.

[10] РОЖДЕСТВЕНСКИЙ Б. Л., ЯНЕНКО Н. Н., Системы квазилинейных уравнений и их приложения к газовой динамике, М., « Наука », 1968.

[11] РЯБЕНЬКИЙ В. С., ФИЛИППОВ А. Ф., Об устойчивости разностных уравнений, М., Гостехиздат, 1956.

[12] САМАРСКИЙ А. А., Введение в теорию разностных схем, М., « Наука », 1971.

[13] САМАРСКИЙ А. А., ГУЛИН А. В., Устойчивость разностных схем, М., « Наука », 1973.

[14] ТИХОНОВ А. Н., САМАРСКИЙ А. А., Уравнения математической физики, М., « Наука », 1972.

[15] BAKHVALOV N., Méthodes numériques, Ed. Mir, 1976.

[16] FORSYTHE G. E., WASOW W. R., Finite Difference Methods for partial Differential Equations, New York-London, 1960.

[17] GODOUNOV S., Equations de la physique mathématique, Ed. Mir, 1973.

[18] GUELFOND A., Calcul des différences finies, Dunod, Paris, 1962.

[19] RICHTMYER R. D., MORTON K. W., Difference methods for initial-value problems, New York [a. o.], Interscience publ., 1967.

[20] YANENKO N., Méthode à pas fractionnaires, Armand Colin, Paris, 1968.

INDEX

Algorithme
de balayage 50, 51
de calcul de la solution du problème de Dirichlet aux différences 271, 274, 279
numériquement instable, exemple 52
Analogue aux différences des formules intégrales de Cauchy et du type de Cauchy 346
Approximation 86, 156
d'une dérivée 73, 74, 95
de l'ordre de h^k 90-93, 156, 158
d'un problème aux limites différentiel 85, 86, 90
par un shéma aux différences 100, 156, 159
Approximation différentielle d'une équation aux différences 214, 355

Babenko-Guelfand, critère de 218, 356
Balayage 50, 51

Condition
de borne de la solution fondamentale 24
de Courant, Friedrichs et Lewy 189, 190, 195, 196
sur la ligne de discontinuité de la solution 242
de Sokhotski-Plemelj 344, 347
Conditionnement, bon conditionnement 39
Conditions aux limites
intrinsèques 347
d'un schéma aux différences, exemples de construction 183, 184
Convergence de l'ordre de h^k 84, 100, 156
Convergence d'un schéma aux différences 82, 84, 100, 155, 156
Critère de bon conditionnement 40-42, 43, 58
d'un problème aux limites aux différences 42, 43, 59
Crière spectral
de borne des puissances de R_h 316
— — — auto-adjoint 317
de stabilité d'un schéma aux différences 124, 126, 127, 199, 200, 205, 316, 317, 324
Critère de stabilité
des schémas aux différences de Samarski 321, 322
des systèmes aux différences pour le problème de Cauchy 113, 210, 211

Décomposition suivant les facteurs physiques 265
Dérivée, remplacement par des rapports aux différences 95
Dérivée partielle, remplacement par des rapports aux différences 166-167

Egalité de Parseval 209, 226
Equation caractéristique 27
Equation aux différences 17, 18
approximation différentielle 214, 355
du second ordre 20, 26
— —, solution fondamentale 28
vitesse de convergence de la solution 71, 100
Equation différentielle sous forme d'une divergence 250
Erreur d'arrondi 52, 65, 99, 131, 283
Espace vectoriel normé 82

Fixation des coefficients aux points intérieurs 215
Fonction définie sur un réseau (discrète) 38, 79, 81-82

développement en série finie de Fourier 224, 233
Formule intégrale de Cauchy, analogue aux différences 346-347
Formule intégrale du type de Cauchy, analogue aux différences 346, 347
Frontière d'un domaine discrétisé 345

Inégalité de l'énergie 229
Invariants de Riemann 304
Itération à pas variable 279

Lissage de la solution d'un problème aux différences 212

Méthode
du balayage 50, 51, 65, 67, 153, 353
—, évaluation des coefficients de balayage 63
—, historique 353
—, justification 55-63
—, variantes 353
de Bélotserkovski-Davydov 265, 355
des caractéristiques 247-248
des conditions aux limites intrinsèques 344, 348
des différences finies 13, 79, 247
de Douglas-Rachford 285
d'étude des problèmes aux différences évolutifs de von Neumann 198
de Newton 154
de relaxation de Fédorenko 288-290
des réseaux 79
de stationnarisation 271
du tir 53, 151
Méthodes itératives, choix de la précision 277
Méthodes variationnelles d'évaluation des valeurs propres 319

Neumann, critère spectral de stabilité du problème de Cauchy aux différences de von 199, 200, 205, 316
Nombre de conditionnement d'un système d'équations linéaires 39
Norme 82, 94, 106-109, 294, 298, 315
énergie 229, 321
d'un opérateur linéaire 116
Noyau absolu, propriétés 340
Noyau de caractéristique a du spectre d'une famille d'opérateurs 339, 356

Opérateur
auto-adjoint 321

Opérateur
projection frontière 348
R_λ 297
—, construction 305
—, critère de R_λ auto-adjoint 318
—, évaluation des normes des puissances 315, 329
—, — des valeurs propres 319-320
Ordre d'une équation aux différences 19, 22
Ordre de précision d'un schéma aux différences 69, 72, 84, 110, 132

Paramètres d'itération
choix optimal 275, 277
ensemble de Tchebycheff 279, 281, 286
numérotation 282, 285
Point du spectre
d'une famille d'opérateurs 326, 327
d'un opérateur 327
Principe
des coefficients fixés 215-217
du maximum 160, 235, 268
Problème
évolutif 292, 324, 330
modèle 187, 224
de suppression de la discontinuité 244, 251
Problème de Cauchy
pour l'équation de la chaleur, schémas aux différences approximants 170, 171
— des ondes, schéma aux différences approximant 205
Problème de Cauchy aux différences 198
analyse de la stabilité 201-208
conditions de stabilité 198, 210, 211
critère spectral de stabilité de von Neumann 199, 200, 205, 316
représentation intégrale de la solution 208, 209
Problème aux limites différentiel
notation symbolique 77-79
solution généralisée 239, 242, 243, 245
Problème aux limites elliptique 266
Problème aux limites général pour des systèmes d'équations aux différences sur un segment maillé, condition de bon conditionnement 47
Problème(s) aux limites aux différences
bien conditionné 39, 55, 62
à coefficients perturbés, évaluation des solutions 55, 57, 58
critères de bon conditionnement 40-43
forme générale 348
notation symbolique 83
procédés de construction 166
stable 159
Processus itératif de Richardson 279

Représentation intégrale de la solution du problème de Cauchy aux différences 208
Réseau 13, 18, 79, 248, 355
Résidu 86

Schéma
d'Adams 145, 147-148
de directions alternées 257, 276, 286
d'Euler 97, 100, 143, 144
Schéma aux différences
absolument stable 238
bilatéral 353
conservatif 250, 252
construction 246, 247, 249
— des conditions aux limites 183
convergent 82, 84, 100, 155, 156, 159
critères de stabilité 113, 123, 124, 126, 190, 196, 198, 199, 210, 211, 299, 301, 302, 317, 318, 321
de décomposition 254, 256, 258, 265, 355
décomposition en sous-systèmes 93, 95
de directions alternées 257, 355
économique 258
pour l'équation de la chaleur 170, 227, 230, 272, 273, 293, 302, 303
pour une équation à coefficients discontinus 353
— aux derivées partielles 155
— intégrale 105
pour l'équation des ondes 205
d'Euler 97, 100, 143, 144
explicite 235, 238
de Godounov 251
implicite 237, 238
inconditionnellement stable 238
instable 76
notation canonique 114, 122, 296
précision 69, 72, 110, 132
prédicteur-correcteur 97, 181, 182, 252, 354
pour le problème des cordes vibrantes 232
— de Dirichlet 171, 195, 196, 266
pour un problème aux limites 85, 86, 100, 156, 159
procédés de construction 95, 97, 152, 156, 157, 166, 172, 246, 247, 249, 254, 354
réduction à la forme canonique 227, 298
rigide 169
de Runge-Kutta 97, 143, 144, 147
stable 98, 99, 113, 159, 160
— par rapport aux erreurs d'arrondi 353
de stationnarisation 272, 273, 356
pour un système d'équations de l'acoustique 304
vérification de la convergence 68, 85, 100, 156
avec viscosité artificielle 247
Série de Fourier pour une fonction discrète 224
Solution fondamentale 23, 30, 345
bornée 30-33
d'une équation aux différences 24, 30, 345
évaluation 30-35
d'un système d'équations aux différences 345
Solution généralisée d'une équation différentielle 239, 242, 245
Spectre d'une famille d'opérateurs
algorithme de calcul 331, 332, 337
définition 326, 327
noyau 338
Stabilité
des algorithmes 52, 65-67, 134, 135, 283-285, 341
des conditions aux limites intrinsèques 348
du problème de Cauchy aux différences, condition nécessaire et suffisante 210, 211
des problèmes non linéaires, procédé d'étude 141, 216
par rapport à la perturbation des données initiales 199
d'un schéma, caractérisation 135
— aux différences 76, 98-100, 113, 159, 160
d'un système aux différences, critère nécessaire 300, 302
— —, — suffisant 123, 210, 299, 306, 318, 322
Suppression de la discontinuité 244, 251

Théorème
de la coïncidence du noyau absolu d'un spectre avec le spectre 340, 356
d'équivalence de Lax 354
de l'inclusion du noyau d'un spectre dans le spectre 339
de l'invariance d'un noyau de caractéristique a 341
du lien entre l'approximation, la stabilité et la convergence 100, 159, 354
de la stabilité d'un système perturbé 132
de la structure du spectre d'une famille d'opérateurs 337
Théorèmes sur les propriétés du noyau absolu 339-341
Transformation de Fourier rapide 355

Viscosité d'approximation 213
Vitesse de convergence de la solution d'une équation aux différences 71, 100, 159